RECPAD'21

27th Portuguese Conference on Pattern Recognition

November 5, 2021
Universidade de Évora

Proceedings

UNIVERSIDADE DE ÉVORA

APRP

Proceedings of RECPAD 2021
Escola de Ciências e Tecnologia
Universidade de Évora
2021

27th Portuguese Conference on Pattern Recognition

https://recpad2021.uevora.pt

Contents

Poster Session 2

Poster Session 3

Preface

This volume collects the papers accepted for RECPAD 2021, the 27th edition of the annual Portuguese Conference on Pattern Recognition, promoted by APRP (Portuguese Association for Pattern Recognition).

RECPAD is a one-day conference that aims to promote the collaboration between the Portuguese scientific community in the fields of **Pattern Recognition, Image Analysis and Processing, Soft Computing and related areas**. Topics include (but not limited to):

- Statistical, structural, syntactic pattern recognition
- Neural networks, machine learning, data mining
- Discrete geometry, algebraic, graph-based techniques for pattern recognition
- Signal analysis, image coding and processing, shape and texture analysis
- Computer vision, robotics, remote sensing
- Document processing, text and graphics recognition, digital libraries
- Speech recognition, music analysis, multimedia systems
- Natural language analysis, information retrieval
- Biometrics, biomedical pattern analysis and information systems
- Special hardware architectures, software packages for pattern recognition

RECPAD 2021 was organized by the Informatics Department of University of Évora and held at Colégio do Espírito Santo (the main building of the University) on November 5th 2021. In this edition **43 papers** were accepted for poster presentation from which 4 were selected for oral presentation. Besides the oral and poster sessions, RECPAD also featured:

- an invited talk by professor Keshav Dahal, from University of the West of Scotland, titled "Multi-objective search with evolving fitness functions for solving scheduling problems"
- prizes for the best oral presentation and best poster sponsored by DECSIS

We would like to express our greatest appreciation to all the authors and members of the scientific and organizing committees which were a key contribution to the success of this conference, this year returning to a physical event after the online event of 2020.

Thank You!

Sponsors & Partners

Committees

Organizing Commitee

Teresa Gonçalves, Universidade de Évora
Luís Rato, Universidade de Évora
Pedro Salgueiro, Universidade de Évora
Francisco Coelho, Universidade de Évora
Miguel Barão, Universidade de Évora
Eduardo Medeiros, Universidade de Évora
Leonel Corado, Universidade de Évora
Rute Veladas, Universidade de Évora

Scientific Committee

Ana Rebelo, Universidade do Porto
Ana Mendonça, Universidade do Porto
António Anjos, Universidade de Évora (ECT)
António Cunha, Universidade de Trás-os-Montes e Alto Douro
Armando Pinho, Universidade de Aveiro
Augusto Silva, Universidade de Aveiro
Beatriz Sousa-Santos, Universidade de Aveiro, IEETA
Bernardete Ribeiro, Universidade de Coimbra
Carlos Ferreira, INESC TEC
Catarina Silva, Universidade de Coimbra, CISUC
César Teixeira, Universidade de Coimbra
Diogo Pratas, Universidade de Aveiro
Fernando Monteiro, Instituto Politécnico de Bragança
Francesco Renna, Universidade do Porto (FCUP), Instituto de Telecomunicações
Hélder P. Oliveira, Universidade do Porto
Hugo Silva, Instituto de Telecomunicações
Inês Domingues, Instituto Superior de Engenharia de Coimbra (ISEC), Centro de Investigação do IPO Porto
Jaime Cardoso, Universidade do Porto
Joana Costa, Institutol Politécnico de Leiria
João Carlos Neves, Instituto de Telecomunicações
João M.F. Rodrigues, Universidade do Algarve

João Moura-Pires, Universidade Nova de Lisboa (FCT)
João Sanches, Universidade de Lisboa (IST), ISR
Joel P. Arrais, Universidade de Coimbra
Jorge Oliveira, Instituto de Telecomunicações
Jorge Santos, ISEP
Jorge Torres, Instituto de Telecomunicações
Jorge S. Marques, Universidade de Lisboa (IST), ISR
Jose Saias, Universidade de Évora (ECT)
José Manuel Fonseca, Universidade Nova de Lisboa (FCT)
Jose Silvestre Silva, Academia Militar
Lio Goncalves, Universidade de Trás-os-Montes e Alto Douro
Luis Teixeira, Universidade do Porto (FEUP), INESC TEC
Luis Rato, Universidade de Évora (ECT)
Luís A. Alexandre, Universidade da Beira Interior, Instituto de Telecomunicações
Mário Antunes, Instituto Politécnico de Leiria, INESC TEC
Miguel Barao, Universidade de Évora (ECT), INESC-ID
Nuno Rodrigues, Instituto Politécnico de Leira (ESTG), Instituto de Telecomunicações
Paulo Salgado, Universidade de Trás-os-Montes e Alto Douro
Pedro Salgueiro, Universidade de Évora (ECT)
Pedro Pina, Universidade de Coimbra
Pedro Jorge, Instituto Superior de Engenharia de Lisboa
Petia Georgieva, Universidade de Aveiro
Rui Neves-Silva, Universidade Nova de Lisboa (FCT)
Teresa Goncalves, Universidade de Évora (ECT)
Verónica Vasconcelos, Instituto Superior de Engenharia de Coimbra (IPC)

Invited Speaker

Keshav Dahal
University of the West of Scotland
https://research-portal.uws.ac.uk/en/persons/keshav-dahal

Biography
Professor Keshav Dahal is a Professor of intelligent systems and the leader of the Artificial Intelligence, Visual Communications and Network (AVCN) Research Centre, University of the West of Scotland, Paisley, U.K. He received his Masters and Ph.D. degrees from the University of Strathclyde, UK. He also worked at the University of Bradford and the University of Strathclyde. His research interests lie in the areas of applied AI, trust and security modelling in distributed systems, Blockchain technology and scheduling/optimization problems. He has been principal investigator or co-investigators on more than 25 externally funded projects, and supervised over 30 PhD and postdoctoral researchers. He has published over 170 papers in his research fields with award winning publications and has sat on organizing/program committees of over 60 international conferences. He is a Senior Member of IEEE.

Title
Multi-objective search with evolving fitness functions for solving scheduling problemss

Abstract
This talk will present some of the development in multi-objective approaches for solving complex scheduling problem. The first part of the talk will investigate multi-objective and weighted single objective approaches to a real world workforce scheduling problem. The computational experiments show that multi-objective genetic algorithms can create solutions whose fitness is close to that of the solution created by the genetic algorithms using weighted sum objectives even though the multi-objective approaches know nothing of the weights. In second part of the talk will discuss the variable fitness function approach to enhance the metaheuristic approaches by evolving weights for each of the multiple objectives. The results show that the variable fitness function approach improves the performance of constructive and variable neighbourhood search approaches on workforce scheduling problem instances.

Conference Program

08:45	Registration
09:15	Opening Ceremony
09:30	Poster Session I
10:30	Coffee Break
11:00	Poster Session II
12:00	Keynote talk by Keshav Dahal
13:00	Lunch Break / APRP meeting
14:30	Poster Session III
15:30	Oral Session
16:30	Coffee Break / historic visit to University of Évora
17:15	Awards and Closing

Hybrid Deep Learning - Hidden Markov Model for Heart Sound Segmentation

Francesco Renna[1]
https://www.inesctec.pt/en/people/francesco-renna
Miguel L. Martins[2]
https://www.inesctec.pt/en/people/miguel-lopes-martins
Miguel Coimbra[2]
https://www.inesctec.pt/en/people/miguel-coimbra

[1] Instituto de Telecomunicações, INESC TEC
Faculdade de Ciências da Universidade do Porto
Porto, Portugal

[2] INESC TEC
Faculdade de Ciências da Universidade do Porto
Porto, Portugal

Abstract

In this paper, we propose a novel algorithm for heart sound segmentation. We introduce a combination of two families of state-of-the-art solutions for such problem, hidden Markov models and deep neural networks, in a single training framework.

The proposed approach was tested with heart sound outperforms current state-of-the-art for the PhysioNet dataset, with an average sensitivity of 93.9% and an average positive predictive value of 94.2% when detecting the boundaries of fundamental heart sounds.

1 Introduction

Each recorded heartbeat on a phonocardiogram (PCG) is composed by two fundamental sounds: the first sound (S1), that is generated by vibrations of the mitral and tricuspid valves at the beginning of the systole, and the second sound (S2), that is generated by the closure of the aortic and pulmonary valve at the beginning of the diastole.

Heart sound segmentation consists on the identification of four fundamental segments in each heart cycle. Analysis of segmented PCG signal allows the detection and localization of extra sound components as well as the access to useful information for morphology analysis of the S1 and S2 waveforms.

Current state-of-the-art solutions for heart sound segmentation can be roughly divided into two classes. The first one leverages statistical models with prior information about the sequential nature of PCG signals, mainly hidden Markov models (HMMs) and hidden semi-Markov models (HSMMs). For both models, each data point composing a heart sound recording is mapped onto a hidden state. In particular, [8] has introduced the use of HSMMs for PCG segmentation, which allow explicit modeling of the time spent in each state, i.e. sojourn time. Other works have improved further the performance of HSMM-based algorithms, by considering a modified Viterbi decoder that addresses boundary conditions [9], or by proposing methods to adapt sojourn time distribution to the specificities of each signal [6].

The second class of segmentation algorithms is based on deep learning. Deep convolutional neural networks (CNNs) have been applied to envelopes extracted from heart sound recordings for heart sound segmentation in [7]. Also, deep sequential models, namely recurrent neural networks (RNNs), have been leveraged to segment PCG signals, by keeping track of the temporal dependencies embedded in the signal [5]. Then, [2] has considered the use of a bidirectional long short-term memory (LSTM) in conjunction with an attention mechanism, which enables identification of the most salient aspect of the signal, thus providing enhanced robustness against noisy and irregular recordings. Finally, [10] has proposed the use of a temporal-framing adaptive network which is trained with a specific transition loss and is able to perform dynamic inference, thus adapting to irregular heart sound behaviors.

Taking inspiration from speech recognition [3], we propose a novel hybrid heart sound segmentation framework which combines the benefits of both HMM-based approaches and more recent deep learning methods. The overall hybrid model is trained end-to-end using a gradient-based approach.

2 Methods

In this section, we describe the proposed approach for heart sound segmentation, providing details about input pre-processing, the definition of the hybrid HMM-deep neural network (DNN) framework, training method, and post-processing.

2.1 Pre-processing

Heart sound recordings are first filtered with a Butterworth filter of order two with pass-band $[25, 400]$ Hz and then processed with the spike removal algorithm presented in [8]. Then, the four envelopes considered in [7, 9] are extracted from the filtered signals (homomoprhic envelope, Hilbert envelope, power spectral density envelope, and wavelet envelope), and normalized to have zero mean and unit variance.

Let $\mathbf{x}_t \in R^4$, for $t = 1, \ldots, T$ be the 4-dimensional signal obtained considering the four envelopes extracted from a given PCG with length T and s_t, for $t = 1, \ldots, T$, the corresponding state label, where $s_t \in \{0, \ldots, L-1\}$ and L represents the total number of possible signal states. We consider $L = 4$, mapping each of the PCG states: S1, systole, S2, and diastole.

2.2 Hybrid model

Our objective is to model the PCG signal via an underlying HMM whose emission probabilities are estimated with a DNN. In contrast, previous work which combined both models [7], our learning strategy *jointly* trains the HMM and the DNN simultaneously using a set of annotated PCG recordings, thus allowing the sequential information contained in the HMM to be used in training for the DNN.

The DNN's feature maps, $\mathbf{o}_t = [\mathbf{x}_{t-F}, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_{t+F-1}] \in R^{4\times 2F}$, are obtained by collecting the signal vectors contained in an observation window for some positive integer F.

Given $s_1^T = [s_1, \ldots, s_T]$, the state sequence, and $\mathbf{o}_1^T = [\mathbf{o}_1, \ldots, \mathbf{o}_T]$, the T-dimensional feature map, we characterize the model using the complete likelihood as follows

$$P(\mathbf{o}_1^T, s_1^T) = p_{s_1} \prod_{t=2}^{T} p_{s_{t-1},s_t} \prod_{t=1}^{T} P(\mathbf{o}_t|s_t), \tag{1}$$

where p_{s_1} is the probability of being in state s_1 at $t = 1$, p_{s_{t-1},s_t} is the transition probability from $(t-1)$ to t for states s_t and s_{t-1}, and $P(\mathbf{o}_t|s_t)$ is the emission probability of the feature map $\mathbf{o}_t$ given the state s_t.

We assume that a DNN parametrized by Θ provided with an observation $\mathbf{o}_t$ outputs $y_{t,s_t,\Theta}(\mathbf{o}_t)$, i.e. the estimates of the probabilities $P(s_t|\mathbf{o}_t)$. Thus, we can express (1) as

$$P(\mathbf{o}_1^T, s_1^T) = p_{s_1} \prod_{t=2}^{T} p_{s_{t-1},s_t} \prod_{t=1}^{T} \frac{P(\mathbf{o}_t)}{P(s_t)} \prod_{t=1}^{T} y_{t,s_t,\Theta}(\mathbf{o}_t). \tag{2}$$

Moreover, the marginal probability of the observation sequence $\mathbf{o}_1^T$ is obtained by summing the joint probability in (1) over all possible state sequences of length T, $s_1^T \in \mathcal{S}$,

$$P(\mathbf{o}_1^T) = \sum_{s_1^T \in \mathcal{S}} p_{s_1} \prod_{t=2}^{T} p_{s_{t-1},s_t} \prod_{t=1}^{T} \frac{P(\mathbf{o}_t)}{P(s_t)} \prod_{t=1}^{T} y_{t,s_t,\Theta}(\mathbf{o}_t). \tag{3}$$

Note that the marginal probability in (3) can be efficiently computed using the forward-backward algorithm [1].

2.3 Training

The training dataset is comprised of a set of pairs $\{(\mathbf{o}_1^T)_n, (s_1^T)_n\}_{n=1}^N$, containing N PCG recordings with the corresponding state sequences. Let $\Psi = [\pi, \Gamma, \Theta]$ be the set all parameters of the model, where π collects the initial state probabilities, and Γ the $L \times L$ transition matrix for the Markov chain. We maximize the maximum mutual information (MMI) criterion [3] using a grandient-based approach:

$$\mathcal{L}(\Psi) = \sum_{n=1}^{N} \log P\left((\mathbf{o}_1^T)_n, (s_1^T)_n\right) - \log P\left((\mathbf{o}_1^T)_n\right), \tag{4}$$

where $P\left((\mathbf{o}_1^T)_n,(s_1^T)_n\right)$ and $P\left((\mathbf{o}_1^T)_n\right)$ are computed using (2) and (3), respectively.

2.4 Inference and post-processing

At inference time, PCG test signals are pre-processed according to the steps described in Section 2.1. Then, the trained model is applied to the features maps obtained from the pre-processed test data and the corresponding outputs $y_{t,s_t,\Theta}(\mathbf{o}_t)$ are used to approximate the emission probabilities $P(\mathbf{o}_t|s_t)$. The output is computed using the Viterbi algorithm [1].

3 Experiments

The proposed hybrid HMM-DNN segmentation algorithm is compared with the HSMM-based method from [9] and the CNN-based approach of [7]. The performance is measured via 10-fold cross-validation over the available heart sound dataset. We ensure that, at each iteration, sounds from patients contained in the test set are not contained in the training set.

We use a simple CNN starting with three blocks of one dimensional convolutions with rectified linear unit (ReLU) activation functions followed by max-pooling layers. A kernel size of 3 with stride of 1 is used throughout all convolutional layers and they stack 8, 16, and 32 filters, respectively. The max-pooling layers have a kernel size and stride of 2. The bottleneck features are passed through a 25% dropout layer and fed to a single hidden dense layer of size 64 using a ReLU activation function. The output layer uses a softmax activation function returning $P(s_t|\mathbf{o}_t) \in R^L$.

The dimension of the DNN input feature maps is $F = 32$. The HMM parameters π and Γ were first estimated via maximum likelihood and then kept fixed while estimating Θ. The latter is attained by maximizing $\mathcal{L}(\Psi)$ using the Adam algorithm [4], with learning rate 10^{-4}. 50 epochs were used per fold, with early stopping implemented by extracting 10% of the training data for validation and retaining the network weights with highest validation loss.

The performance of the considered segmentation algorithms in determining the position of the fundamental heart sounds S1 and S2 is evaluated in terms of their sensitivity (S) and positive predictive value (P_+). Such metrics are computed according to the description in [8], where true positives are counted when the mismatch between the center of a sound in the estimated sequence and ground truth sequence is lower than 60 ms. All performance metrics are computed for each recording in the test set and then averaged over the test set. Finally, the values corresponding to the different 10 test subsets are reported.

The heart sounds used for the experiments were taken from the dataset made publicly available for the PhysioNet/CinC challenge 2016. In particular, we considered 792 heart sounds recorded from 135 patients in different clinical and non-clinical environments.[1] From those, 181 sounds are collected from patients with pathological heart lesions (most commonly mitral valve prolapse), as assessed by echocardiography. The remaining 246 sounds are collected from healthy patients. Sound recordings are sampled at 1 kHz. The annotations provided with the dataset are obtained from the analysis of synchronous ECG recordings.

In Fig. 1 are reported the values of sensitivity (S) and positive predictive value (P_+) for the proposed method and the algorithms in [9] and [7]. It is possible to note that the proposed methods outperforms the segmentation algorithms considered for comparison both in terms of sensitivity and positive predictive value. These results can be explained by the enhanced capability of the proposed hybrid model in embedding explicitly domain knowledge regarding the sequential nature of the PCG signal when compared to the application of a CNN classifier followed by Viterbi decoding.

4 Conclusion

In this work, a hybrid model-based/data-driven approach for heart sound segmentation was presented. The proposed framework consists in the joint training of a HMM, able to explicitly embed sequential information about heart sound states, and a highly discriminative DNN, via a mutual information maximization criterion.

Such approach displays superior segmentation performance compared to current CNN architectures and HSMM approaches, motivated by the high flexibility of the hybrid model in striking a balance between explicit sequential modeling and discrminiative power.

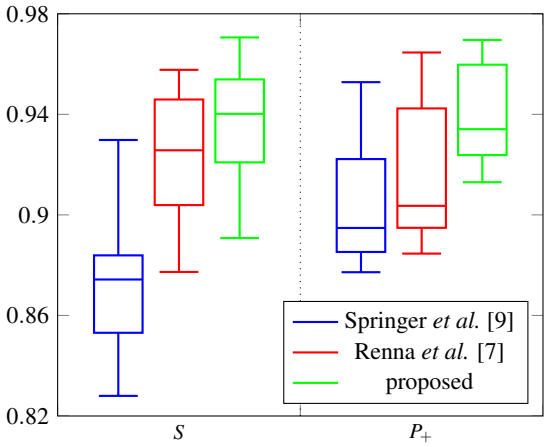


Figure 1: Sensitivity (S) and positive predictive value (P_+) of the HSMM-based method in [9] (blue, left), CNN-based method in [7] (red, center), and proposed method (green, right).

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

References

[1] C.M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] Tharindu Fernando, Houman Ghaemmaghami, Simon Denman, Sridha Sridharan, Nayyar Hussain, and Clinton Fookes. Heart sound segmentation using bidirectional lstms with attention. *IEEE Journal of Biomedical and Health Informatics*, 24(6):1601–1609, 2020. doi: 10.1109/JBHI.2019.2949516.

[3] Lior Fritz and David Burshtein. Simplified end-to-end MMI training and voting for ASR. *arXiv:1703.10356*, 2017.

[4] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[5] E. Messner, M. Zöhrer, and F. Pernkopf. Heart sound segmentation: An event detection approach using deep recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 65(9):1964–1974, 2018. doi: 10.1109/TBME.2018.2843258.

[6] J. Oliveira, F. Renna, and M. T. Coimbra. Adaptive sojourn time HSMM for heart sound segmentation. *IEEE Journal of Biomedical and Health Informatics*, 23(2):642–649, 2019.

[7] F. Renna, J. H. Oliveira, and M. T. Coimbra. Deep convolutional neural networks for heart sound segmentation. *IEEE Journal of Biomedical and Health Informatics*, 23(6):2435–2445, 2019.

[8] SE Schmidt, Claus Holst-Hansen, Claus Graff, Egon Toft, and Johannes J Struijk. Segmentation of heart sound recordings by a duration-dependent hidden Markov model. *Physiological measurement*, 31(4):513–529, 2010.

[9] David B Springer, Lionel Tarassenko, and Gari D Clifford. Logistic regression-HSMM-based heart sound segmentation. *IEEE Transactions on Biomedical Engineering*, 63(4):822–832, 2016.

[10] Xingyao Wang, Chengyu Liu, Yuwen Li, Xianghong Cheng, Jianqing Li, and Gari D. Clifford. Temporal-framing adaptive network for heart sound segmentation without prior knowledge of state duration. *IEEE Transactions on Biomedical Engineering*, 68(2):650–663, 2021. doi: 10.1109/TBME.2020.3010241.

[1] The sounds are available online at https://PhysioNet.org/physiotools/hss/.

Using Knowledge Distillation to Interpret Credit Score Modeling

Tiago Faria
tiagofaria@student.dei.uc.pt
Catarina Silva
catarina@dei.uc.pt
Bernardete Ribeiro
bribeiro@dei.uc.pt

Universidade de Coimbra
CISUC - Centro de Informática e Sistemas
FCTUC-DEI - Departamento de Engenharia Informática
Coimbra, Portugal

Abstract

In the last decade many accurate decision support systems have been constructed as black boxes. However, applicability in several critical applications, e.g. public policy, security/safety systems, health diagnosis and fraud detection, has been faced with some hurdles due to lack of model interpretability. In this work we present knowledge distillation as a stepping stone to achieve model interpretability by interpretable models mimic more complex ones such as deep neural nets. We show that there's a possibility for less complex but interpretable models to mimic deep neural nets, by giving transforming classification problems into a regression problem.

1 Introduction

The financial industry is highly regulated and in the case of loan issuers, laws around the world, e.g. the European Union General Data Protection Regulation (EU GDPR), start to determine that in a not far away future, financial institutions must effectively show that the decisions they take are fair. The systems implemented in the financial sector are usually black-box models, highly capable of achieving their goal with high performance. The problem with black-box models is, although they are usually very capable, their decision processes are not clear and also prone to bias. Thus, one significant challenge of using AI-based systems that, for instance predict credit scores, is that there is no underlying interpretability infrastructure that can provide reason code to borrowers, e.g.,when a credit is denied.In this work we propose a method that uses knowledge transfer from deep models to decision-tree models in an attempt to understand the decision patterns in financial applications.

2 Background

Machine learning critical decision-making is a relatively recent topic. As humans get assisted or even replaced by intelligent models, existing legislation becomes obsolete and data regulation is often ineffective. Hence, new regulations like the European Union General Data Protection Regulation (EU GDPR), which includes article 22 on automated decision making are establishing the need for interpretability in the sector. Although still in debate, the GDPRs article 22 clauses on automated individual decision-making have introduced the right to explanation [1] for all individuals to obtain "meaningful explanations of the logic involved" while being targets of automated decision-making algorithms. As safeguard for the companies implementing these models, as well as for the subjects that are targeted, interpretability starts to become essential in the transition to fully digital automated services. In fact, some companies are starting to learn the problems of black-box models in their services [2]. Knowledge Distillation was first introduced in 2015 [3] and is a generalization of **Model Compression** [4]. **Model Compression** consists on the transfer of learned knowledge of as lower, larger and better performing model onto a smaller, faster. Caruana et al. [4] achieves this by matching the logits of the smaller model to the logits of a cumbersome model. This means that the smaller model will approximate the behaviour of the more complex one by training on big ammounts of pseudo-data (logits) which in turn will get better results than the same model trained on real data given there's more information stored on logits than there is on hard labels. **Knowledge Distillation** is a variant of this approach proposed by Hinton et al. [3] which uses the last layer's soft probabilities instead of logits as targets for training a smaller deep neural net student model.

3 Proposed Approach

Although knowledge distillation is not a new topic, we believe that there's more to do with it when it comes to interpretability, the interaction between classes can be a good resource to explain how a model came up with a certain decision. Some work in distilling knowledge to interpretable models has been done by Che et al.[5], but the models used were GAM's and splines, which don't have a great visualization, decision-trees are very easy to visualize and better at capturing feature interaction.

We propose distilling knowledge from a deep neural net to a decision-tree by training a deep neural net model using a dataset **{X,y}** which is often called **Teacher** (this could also be a previously trained model), we then use the Teacher's softmax layer output **y'** as targets for a decision-tree regressor which we call the Student. While the teacher has learned classification, the student will simply try to match the teacher. In theory if we can achieve a perfect score in the student, we get a surrogate model that is easily interpretable.

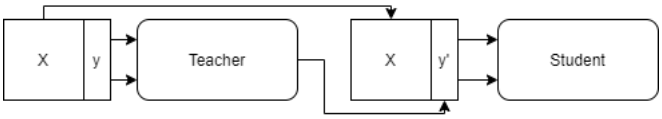


Figure 1: Knowledge distillation process

Methods to interpret trees can be more intuitively easy to come up with and explore. We believe that the tree structure is ideal for capturing interaction between features in data, visualization of decision-trees is also human-friendly making them better for explanation and interpretation. Figure 1 depicts the knowledge distillation process for the proposed approach.

We propose distilling knowledge from a deep neural net to a decision-tree by matching logits (scores before the last softmax layer), we do this by using these logits as targets to train a decision-tree for regression. This decision-tree should in theory mimic the way the deep neural net makes its decisions. It should capture not only the good parts but the bad parts. This tree can then be evaluated for interpretability.

4 Experimental Setup

4.1 Dataset

The data used on this project was kindly provided and given permission to work on by Jörg Osterrieder and Branka Misheva as part of **COST**. The dataset is comprised of 113937 instances with each consisting of a group of 80 descriptive attributes that characterize the outcome of an individuals loan given by Prosper. Prosper Marketplace, Inc. is an american company in the peer-to-peer lending industry. Data was cleared of all null values and unnecessary columns and consisting of 106290 instances and a total of 59 attribute columns at the end of the clearing process. LoanStatus represents the target for classification that initially consisted of 11 classes.

The instances classified as Current were dropped as they had no real value on the training, predicting of any of the models since we don't know what the final outcome was. The rest of the classes were grouped up with all Past Due becoming a new Problematic class; Charged off and Cancelled grouped up with Defaulted, and FinalPaymentInProgress coupled with Completed for the sake of keeping as much data as possible, as so we are left with a 3 class problem with the classes, Defaulted, Problematic and Completed. This leaves us with a dataset comprised of 49724 entries.

Table 1: Initial classes and their distribution

Class Name	Number of Instances
Current	56566
Completed	33530
Defaulted	3289
Past Due (aggregated 1-120+ days)	2067
FinalPaymentInProgress	205
Charged-off.	10632
Cancelled	1

4.2 Methodology

A deep neural net was trained in classifying the 3 classes, using a 70% of the total dataset for training, leaving 30% for testing, after hyper-parameter optimization, the best neural network was chosen to be the teacher. After training we pass the full training dataset **X** once again through the same neural network obtaining a list of probabilities vectors **y'** size $n \cdot c$ where n is the number of instances of the training dataset and c is the number of classes in the classification problem, in our case $c = 3$. After we obtain the new set **{X,y'}** we use it to train a decision tree regressor, which we call the student. We than compare the student and teacher for similarity in the predictive power by looking at the respective scores for the classes. A second decision-tree is trained in order to validate the differences between training a model using knowledge distillation and training a model on ground-truth labels.

4.3 Evaluation Metrics

Tests were done using basic metrics for model evaluation that tell us how well a model is performing, which are then to be compared between the "teacher" and "student" models. In cases where the weight of false positives and false negatives have different cost or in unbalanced datasets its better to use metrics that difference into account, as such for our problem we look to F1-score as being a more important metric than accuracy. If the values of precision and recall across all three classes classified by the student model is somewhat similar or better than the teacher model we can presume that its reliable to replace the teacher with this model.

5 Results and Analysis

After both models were trained we used the test portion of the dataset to assess the validity of the method, obtaining the following results.

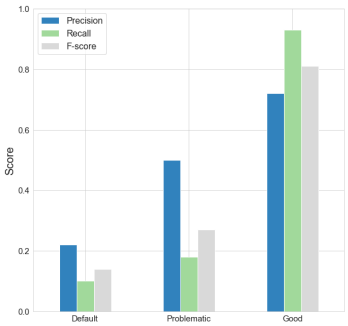


Figure 2: Teacher results

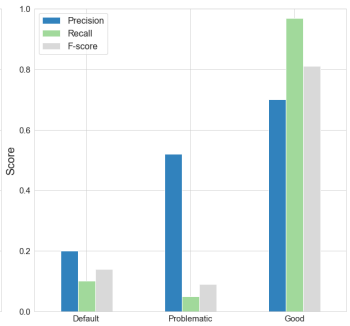


Figure 3: Student results

As we can see, on figures 2 and 3 we see extremely similar scores on both student and teacher, as it was said before, the student is meant to copy it, this means that it will also try to capture the worse parts as we can see from the problematic part. While if we look at the results for the ground-truth (see figure 4) model we see that these look quite different. This happens because we've given more information to the the model trought the labels by transforming a classification problem into a regression problem, giving the model more "in-between" values that it can guide himself with making it easier to achieve higher accuracies, this can be seen as a form of pre-processing. On the other hand, a model trained with hard-labels sees every instance of the same class as exactly the same, not taking into account the possible similarities or relationship with other classes, this makes the results differ based on the structure and parameters of the model itself.

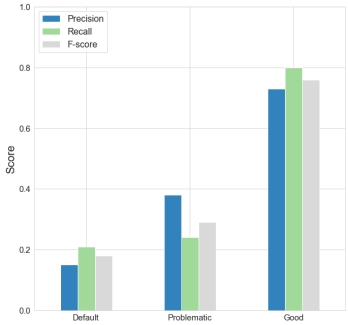


Figure 4: Model trained on ground-truth results

Table 2: Scores for the three models

Class	Model	Precision	Recall	F_1	Acc
Default	Teacher	0.22	0.10	0.14	0.68
	Student	0.20	0.10	0.14	0.68
	Ground-truth	0.15	0.21	0.18	0.62
Problematic	Teacher	0.50	0.18	0.27	0.68
	Student	0.52	0.05	0.09	0.68
	Ground-truth	0.38	0.24	0.29	0.62
Good	Teacher	0.72	0.93	0.81	0.68
	Student	0.70	0.97	0.81	0.68
	Ground-truth	0.73	0.80	0.76	0.62

6 Conclusions and Future Work

In this work we show the potential of using knowledge distillation to improve a less complex model's accuracy, in our experiment we achieve extremely similar scores on both teacher and student. This leads us to think that if we improve the teacher's prediction accuracy for the minority class, will have an interpretable model, in our case a decision-tree that performs better than the same model trained on ground-truth labels.

To improve performance on the teacher as it is a very complex problem with a very particular emphasis on the fact that it is very imbalanced, future work would be on improving the pipeline to get better results on the teacher, for example by creating an ensemble of neural models, or by distilling a teacher onto another neural net which has shown to be effective.

Acknowledgements

This work acknowledges research support by COST Action "Fintech and Artificial Intelligence in Finance - Towards a transparent financial industry" (FinAI) CA19130 (`https//fin-ai.eu/`).

References

[1] Bryce Goodman and Seth Flaxman. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine*, 38(3):50–57, 2017. doi: 10.1609/aimag.v38i3.2741.

[2] Taylor Telford. Apple Card algorithm sparks gender bias allegations against Goldman Sachs, 11 2019. URL `https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-\-allegations-against-goldman-sachs/`.

[3] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[4] Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. volume 2006, pages 535–541, 08 2006. doi: 10.1145/1150402.1150464.

[5] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Distilling knowledge from deep networks with applications to healthcare domain, 2015.

xECG: Using Interpretability to Understand Deep ECG Biometrics

João Ribeiro Pinto[1,2]
joao.t.pinto@inesctec.pt

Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt

INESC TEC
Porto, Portugal

Faculdade de Engenharia, Universidade do Porto
Porto, Portugal

Abstract

The literature on electrocardiogram (ECG) biometrics frequently indicates that the QRS complex is the most important part of this signal. Some go further and argue that just the QRS is enough for accurate ECG biometric identification. To verify this claim, this work uses interpretability tools to analyse how a state-of-the-art deep learning model uses each part of the ECG signal to reach identity decisions under different signal quality conditions and population sizes. Results indicate that the QRS is indeed the most relevant part of the ECG for identification, but its relative importance is smaller in more realistic scenarios. Such insights could be used as regularisation to avoid excessive focus on the QRS complex and thus achieve more robust models.

1 Introduction

The ECG measures the conduction of electrical potentials across the heart's muscle that controls its contraction and relaxation. The cyclical repetition of depolarisation and repolarisation of cells on the heart's atria and ventricles turns the ECG into a repetition of easily recognisable heartbeats. Each heartbeat is composed of a P wave, a QRS complex, and a T wave and, since its morphology depends on the heart structure, it carries important identity information [9].

ECG signals have been successfully used for several automatic pattern recognition tasks, including biometric recognition [9]. Once focused on clean medical signals (*on-the-person* signals), the field of ECG biometrics has decisively evolved towards signals acquired in more realistic biometric scenarios (*off-the-person*). In such scenarios, where noise and variability are prevalent, deep learning methods [3, 5, 7, 10] have enabled higher accuracy and robustness.

However, while traditional methods based on fiducial features are fairly transparent, it is not easy to understand what information deep learning models use to reach a decision. Considering the well-known stability and uniqueness of the QRS complex [4], one can assume that models will focus primarily on this landmark, but this assumption may be incorrect. While pioneer methods used only information from the QRS, this practice has become increasingly uncommon. This may indicate that the QRS may not be sufficient for recognising identity in some scenarios.

These doubts on the role played by the signal waveforms in modern ECG biometric recognition can be understood through interpretability tools. These tools have been developed following the growing awareness of the paramount importance of transparency in artificial intelligence. They enable the analysis of the inner workings of machine learning models applied to diverse pattern recognition tasks [2]. Hence, instead of returning to simpler algorithms such as decision trees (sacrificing performance in favour of transparency), interpretability allows us to obtain sophisticated models which are both highly accurate and transparent.

Hence, the work described in this paper[1] aims to understand the behaviour of deep ECG biometrics through interpretability [8]. A state-of-the-art model for ECG identification [7, 10] was trained on diverse scenarios, with on-the-person and off-the-person data from growing sets of identities, emulating increasingly challenging conditions. Then, interpretability tools were used to evaluate which parts of the ECG signal are most relevant to the model's decisions.

2 Methodology

This study consisted of (1) training an identification model in diverse scenarios, (2) inferring for test samples and analysing identification accuracy, (3) obtaining decision explanations for each test sample, and (4) visualising and analysing the results. Details on the methods are presented below.

[1]Code and additional results available at `https://github.com/jtrpinto/xECG`.

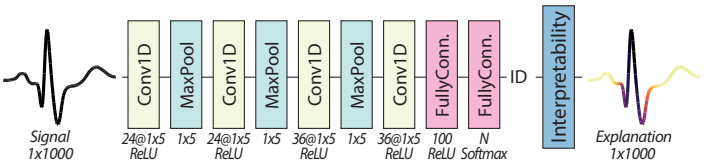


Figure 1: Schema of the methodology: the CNN model is trained to offer identity decisions, which are then explained by interpretability tools.

Model The model used for identification (see Fig. 1) is the one proposed by Pinto *et al.* [7, 10], which has achieved state-of-the-art performance in both identification and identity verification. It receives a five-second ECG segment as input and delivers probabilities for each identity on the training set. The model is an end-to-end convolutional neural network (CNN) with four convolutional layers (two with 24 1×5 filters followed by two with 36 filters) interposed with three 1×5 max-pooling layers. To deliver N identity probabilities, the model concludes with two fully-connected layers (100 and N neurons) and softmax activation.

Interpretability Tools To quantify the relative importance of ECG waveforms on the decisions of the trained identification models, four interpretability tools, from the Captum toolbox for PyTorch, were used:
- *Occlusion* [14] works by hiding parts of the input and measuring the corresponding change in the model's outputs. More relevant input regions will correspond to larger changes in the output;
- *Saliency* [12] uses backpropagation to compute the gradients of target class scores w.r.t. the input. The resulting class score derivatives are rearranged into a saliency map, which assigns higher relevance to input regions that correspond to higher gradients;
- *Gradient SHAP* [6] considers the explanations of a model's predictions as models themselves. Explanation models are simplified and interpretable approximations of the sophisticated models that generate them. SHapley Addictive exPlanation (SHAP) values are computed and denote how much each input region raises the probability for a given class;
- *DeepLIFT* (Deep Learning Important FeaTures) [11] uses backpropagation to trace output contributions to the responsible regions of the input. It compares differences in inputs and outputs based on a baseline input and assigns contribution scores to each neuron of the model.

Visualisation Just as in image interpretability, signal explanations should be visualised in a way that illustrates input morphology and local relevance simply and clearly. Thus, this work proposes a visualisation methodology based on multicoloured line plots. The colour of each part of the signal depends on its relevance for the model's decision: less relevant parts are in light yellow and most relevant parts are in darker purple.

3 Experimental Setup

This work used data from the PTB ECG database [1] and the University of Toronto ECG database (UofTDB) [13]. The PTB includes on-the-person signals from 290 subjects at rest. The UofTDB includes off-the-person recordings from 1019 volunteers during up to six sessions and five postures. Five-second segments were blindly extracted from the recordings. Half of the segments from each identity were used during training and the remaining were used for testing.

Growing subsets of identities (subjects) are considered to emulate increasing challenging scenarios. Within each database, the N first identities are selected, with $N \in \{2, 5, 10, 20, 50, 100, 200, 500, 1019\}$. Identities #1 and #2 are thus the only ones present in all subsets. To take full advantage of the PTB dataset, the entire set of 290 identities was used instead of the 200 identities subset.

Table 1: Identification accuracy results (%) on the test data.

Database	Number of Identities								
	2	5	10	20	50	100	200[1]	500	1019
PTB	100.0	100.0	99.63	99.50	98.92	98.76	97.73	-	-
UofTDB	100.0	97.26	98.30	95.46	93.86	91.16	89.70	91.20	91.45

[1]For PTB, this column corresponds to the entire set of 290 subjects.

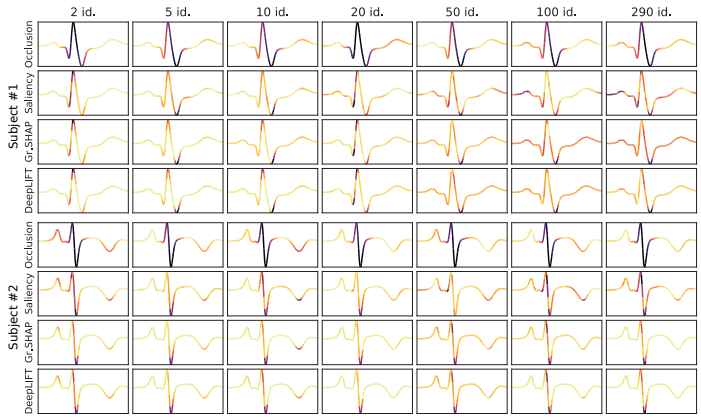


Figure 2: Model explanations on the PTB database.

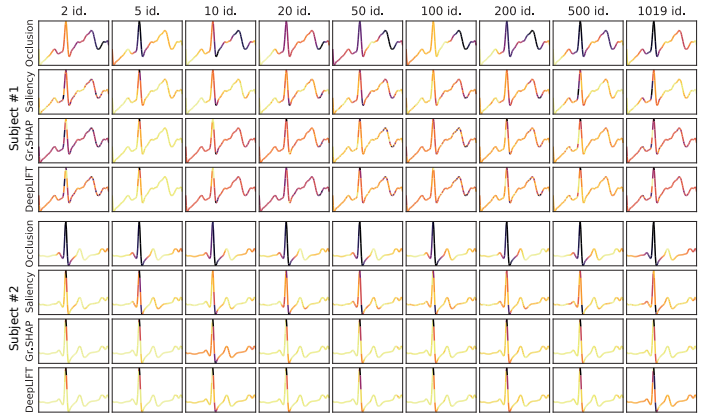


Figure 3: Model explanations on the UofTDB database.

4 Results and Discussion

After training in the aforementioned diverse scenarios, the models' identification accuracy was evaluated (see Table 1). The results follow the expected trends: the model is significantly more accurate in smaller populations for both databases. However, increasing the number of subjects leads to faster performance decay when using off-the-person data. This is evidence of the yet unsolved challenges of increased noise and variability in off-the-person ECG biometrics.

The explanations obtained using the interpretability tools were combined into average heartbeat relevance maps for each subject in each scenario (see Fig. 2 and Fig. 3). One can observe that, with cleaner signals from PTB, acquired in medical conditions, the models focus mainly on the QRS complex. However, as the set of identities grows, the models start to use some information about other waveforms for their decisions. Nevertheless, with these signals, the stability and information of the QRS complex are enough for the model to largely ignore the rest of the signal.

With more realistic signals from UofTDB, the QRS, although important, shares the relevance for the decision more evenly with the other ECG waveforms. These results show that, while the models still prefer the QRS complex, the challenging scenarios of noisy signals and larger populations lead them to look for broader sources of identity information and take advantage of the other ECG waveforms. Another interesting result is that Occlusion generally attributes greater relevance to the QRS than other interpretability tools. This occurs even in the most challenging scenarios. This shows that the deep learning model, not unlike traditional methods, may be learning to locate the different ECG waveforms using the QRS as a reference landmark: when occluded, the effect on the output is larger than with other interpretability methods.

5 Conclusion

This work used interpretability tools to study how state-of-the-art deep biometric models use ECG signals to distinguish people. In general, this study found that the literature is correct in hailing the QRS complex as the most important part of the ECG for biometrics. However, this is more evident in less challenging scenarios. Considering more realistic scenarios, with larger populations and lower quality signals, the relevance is more evenly shared between the ECG waveforms. This indicates that, although still important, the QRS is no longer enough for robust identification.

Thus, one should avoid placing the entire burden of identification upon any single part of the ECG, including the QRS. Every part of the signal carries information that is important in realistic scenarios. These insights could be used as regularisation to promote behaviours that lead towards more accurate and robust models. Additionally, further efforts should be devoted to extending this study into more thorough and objective analyses of ECG waveform relevance.

Acknowledgements

This work was partially funded by the Project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership and the Ph.D. grant "SFRH/BD/137720/2018".

References

[1] R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. *Biomedizinische Technik*, 40(1), 1995.

[2] D. Carvalho, E. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 2019.

[3] M. Hammad, P. Pławiak, K. Wang, and U. R. Acharya. ResNet-Attention model for human authentication using ECG signals. *Expert Systems*, 2020.

[4] R. Hoekema, G. J. H. Uijen, and A. van Oosterom. Geometrical aspects of the interindividual variability of multilead ECG recordings. *IEEE-TBME*, 48(5):551–559, 2001.

[5] R. D. Labati, E. Muñoz, V. Piuri, R. Sassi, and F. Scotti. Deep-ECG: Convolutional Neural Networks for ECG biometric recognition. *Pattern Recognition Letters*, 126:78–85, 2018.

[6] S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In *NeurIPS*, pages 4765–4774, 2017.

[7] J. R. Pinto and J. S. Cardoso. An End-to-End Convolutional Neural Network for ECG-Based Biometric Authentication. In *BTAS*, 2019.

[8] J. R. Pinto and J. S. Cardoso. Explaining ECG Biometrics: Is It All In The QRS? In *BIOSIG*, 2020.

[9] J. R. Pinto, J. S. Cardoso, and A. Lourenço. Evolution, Current Challenges, and Future Possibilities in ECG Biometrics. *IEEE Access*, 6:34746–34776, 2018.

[10] J. R. Pinto, J. S. Cardoso, and A. Lourenço. Deep Neural Networks For Biometric Identification Based On Non-Intrusive ECG Acquisitions. In *The Biometric Computing: Recognition and Registration*, chapter 11, pages 217–234. CRC Press, 2019.

[11] A. Shrikumar, P. Greenside, and A. Kundaje. Learning Important Features through Propagating Activation Differences. In *ICML*, pages 3145–3153, 2017.

[12] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *ICLR*, 2014.

[13] S. Wahabi, S. Pouryayevali, S. Hari, and D. Hatzinakos. On Evaluating ECG Biometric Systems: Session-Dependence and Body Posture. *IEEE-TIFS*, 9(11):2002–2013, 2014.

[14] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In *ECCV*, pages 818–833, 2014.

MOSNet: A light-weight Moving Object Segmentation Network for Autonomous Driving

Gopi Krishna Erabati
gopi.erabati@isr.uc.pt
Helder Araujo
helder@isr.uc.pt
Institute of Systems and Robotics
University of Coimbra, Portugal

Abstract

The ability to segment moving objects like cars is a very crucial element of visual perception system of autonomous vehicles for safe manoeuvrability of vehicles. In this paper, we aim to propose a light-weight Moving Object Segmentation Network (MOSNet) which adapts a two-stream architecture to extract appearance and motion features from RGB images and optical flow respectively. The extracted features are fused with the help of a fusion transformer, a Feature Pyramid Network (FPN) head is used to combine feature maps at various scales and further they are bilinearly upsampled to get back to the original dimension of the input which produces per-pixel class. The network is trained and tested on publicly available KITTI MOD dataset. It is shown that the proposed architecture achieves the Intersection over Union (IoU) of 48.89 % for moving objects and also runs at 50 fps on a RTX 2080 Ti GPU using a ShuffleNetV2 backbone.

1 Introduction

The field of autonomous driving has been progressing rapidly by leveraging deep learning. Visual perception is a key element of autonomous driving which semantically reasons the scene for safe control and maneuverability of the vehicle. It is very critical for a vehicle to know the moving objects in a very complex dynamic traffic environment for optimal planning and control of the vehicle.

In this work, we propose a two-stream architecture MOSNet to segment moving objects in autonomous driving scenarios. Specifically, we extract appearance features from RGB images and motion features from optical flow using ShuffleNetV2 [3] backbone. The extracted multi-level appearance and motion features are fused together with attention mechanism of transformers [6] to capture global appearance and motion cues. A FPN [2] is used to combine multi-scale features. The fused multi-level features are upsampled by convolutions and bilinear upsampling until it reaches a stride of 4 and then all multi-level features are finally summed and transformed to pixelwise output similar to [1].

The contributions of this work are: 1) We present an architecture MOSNet to segment moving objects using appearance and motion features, fusing them with a transformer and bilinearly upsampling to produce pixelwise output. 2) The network is trained and inferred on publicly available KITTI MOD dataset [5]. An IoU of 48.89 % is achieved for moving objects. The network runs in real-time with an inference speed of 50 fps on a RTX 2080 Ti GPU which is very critical for time-constrained applications like autonomous driving.

The paper is organized as follows: Section II provides the details about the architecture, Section III shows experimental results and finally Section IV provides the conclusion.

2 MOSNet

The Moving Object Segmentation (MOSNet) architecture is shown in Fig. 1. A light-weight network pretrained ShuffleNetV2 is used as a backbone feature extraction network to extract both appearance (ShuffleNetV2_RGB) and motion (ShuffleNetV2_OF) semantic features from RGB images and optical flow respectively. The key idea behind ShuffleNet is to use grouped convolutions and channel shuffling to make the network computationally efficient. ShuffleNetV2 architecture consist of a four stage network which provides four different feature scales.

The multi-scale RGB and optical flow feature maps are fused together in the fusion block using an attention based transformer [6] network. The idea here is to exploit self-attention mechanism of transformers to capture global appearance and motion cues from the multi-scale feature maps. Let the intermediate feature map be of dimension $H \times W \times C$, the RGB and OF feature maps are stacked together to form a sequence of dimension $(2 * H * W) \times C$. The input sequence along with positional encoding is fed to the transformer. The transformer produces an output of same dimension as input, so the output is then reshaped to 2 feature maps of dimension $H \times W \times C$ and fused with the corresponding feature maps by element-wise summation. The same fusion strategy is applied for all multi-scale feature maps. However, self-attention on high dimensional feature maps is computationally expensive, so we downsample the feature maps to a fixed resolution of $H = W = 8$ before passing through the transformer and then bilienarly upsample the outputs back to their original resolution. A standard FPN [2] with 128 output channels is used to combine low resolution strong semantic feature maps with high resolution weak semantic feature maps using bilinear upsampling and lateral connections to improve the accuracy of the network.

A FPN head network is used to merge the features from all levels of FPN pyramid into a single output. The low resolution (scale 1/32) feature map is passed through three upsmapling stages to obtain a feature map at 1/4 scale, where each upsampling stage consist of 3×3 convolution, normalization, ReLU activation and $2\times$ bilinear upsampling. A similar procedure is followed for other feature maps at 1/16, 1/8, 1/4 scales with less number of upsampling stages to obtain feature maps at 1/4 scale. Finally, the four feature maps at 1/4 scale are element-wise summed, passed through a $4\times$ bilinear upsmapling layer and a softmax layer to obtain per-pixel class labels (here classes=2; static and moving) at original resolution.

3 Results and Discussion

3.1 Dataset and Implementation Details

The MOSNet architecture is trained and tested on KITTI MOD [5] dataset. The dataset consists of moving object segmentation masks for six sequences of the KITTI raw dataset which accounts for 1750 frames. The moving cars are segmented in the dataset, which accounts for 2383 moving cars. The training split consist of 1300 frames. We only segment 'car' class as the dataset contains moving labels only for 'car' class.

The model is trained end-to-end using weighted cross-entropy loss to account for class imbalance. The class weight is calculated as $w_{class} = \frac{1}{\ln(c+p_{class})}$. Adam optimizer with a learning rate of $1e^{-5}$ along with a weight decay of $5e^{-4}$ is used to avoid overfitting. The model is trained with a batch size of 16 on NVIDIA RTX 2080 Ti GPU for 150 epochs, which accounts for 15 hours of training. The evaluation metrics used in segmentation are Intersection over Union (IoU), precision, recall and F-score.

3.2 Results

The quantitative moving object segmentation results of MOSNet are shown in Tab. 1. MODNet [5] is jointly trained for object detection and segmentation, but we only compare with a separate training strategy for segmentation of MODNet which relates to our work. The MODNet uses a FCN-8s decoder network to upsample the feature maps without a FPN to combine multi-scale feature maps, and they also fuse the features only by element-wise summation. In our network we exploit self-attention of transformers to robustly fuse the features which improves the moving IoU score of our network and we also employ FPN network to combine multi-scale feature maps, which helps to improve the accuracy. We achieve over 10 % gain in moving IoU as compared with MODNet. FuseModNet uses LiDAR data to improve the results. We not only obtained a competitive moving IoU compared to FuseMODNet but also our network achieved an inference speed of 50 fps which is double that of FuseMODNet.

We also employed a light-weight MobileNetV3 as a backbone and a FCN-8s decoder with a Squeeze and Excite kind of strategy to improve

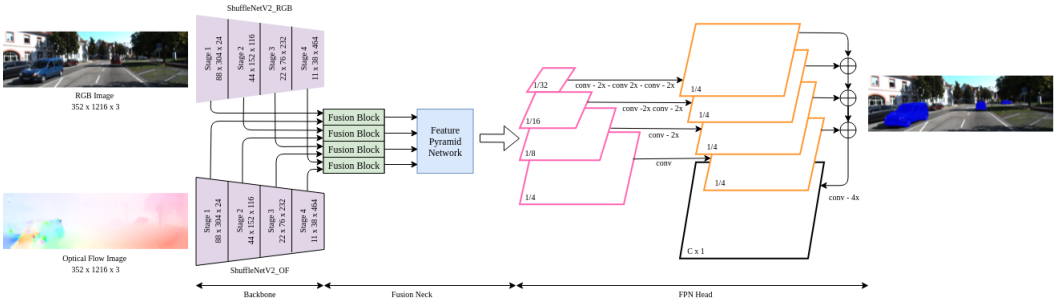


Figure 1: MOSNet Architecture

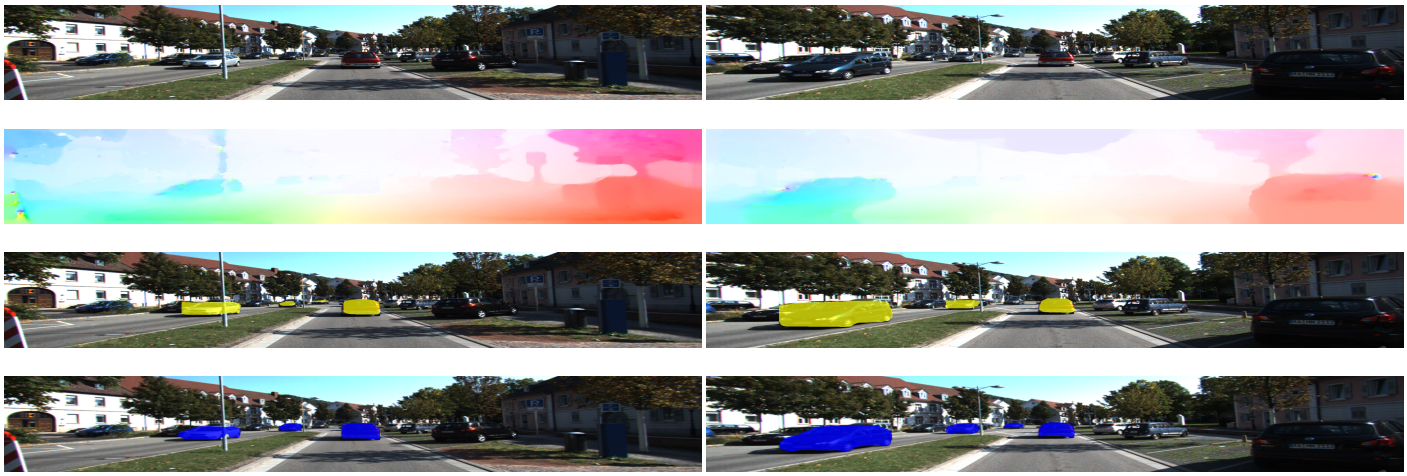


Figure 2: Qualitative results of MOSNet on two samples of KITTI MOD test data. (1st row) Input RGB Images. (2nd row) Input Optical Flow. (3rd row) Ground truth moving object segmentation shown in yellow. (4th row) Output moving object segmentation of MOSNet shown in blue.

Method	Modality	Precision	Recall	F-score	Moving IoU
MODNet [5]	Camera	44.34	69.84	54.25	37.22
FuseMODNet [4]	Camera + LiDAR	-	-	-	51.46
MOSNet (ShuffleNetV2 + FPN)[Ours]	Camera	59.55	78.65	67.78	48.89
MOSNet (MobileNetV3 + FCN)[Ours]	Camera	65.09	62.51	63.77	47.81

Table 1: Quantitative results of our proposed Moving Object Segmentation Network on KITTI MOD dataset

the decoder. Using MobileNetV3 backbone we achieved a moving IoU of 47.81 and the inference speed was approximately 30 fps. The MOSNet with ShuffleNetV2 as a backbone with FPN head was able to provide a good trade-off between accuracy and inference speed which is very crucial for real-time applications like autonomous driving.

The qualitative results of our approach on KITTI MOD test data are shown in Fig. 2. The moving cars on the road are segmented by MOSNet and are shown in blue mask in bottom row of the Fig. 2. A video demonstrating the performance our network on KITTI MOD test dataset is available at [1].

4 Conclusion

In this work, a two-stream Moving Object Segmentation Network (MOSNet) is proposed to segment the moving objects. Specifically, we extract appearance and motion features from RGB and optical flow images using pretrained ShuffleNetV2 backbone, use a self-attention based transformer network to capture global features, fuse both the features with element-wise summation and use a FPN to combine strong and weak semantic features to obtain per-pixel class as output. We trained and tested the network on publicly available KITTI MOD dataset. We achieved 48.89 % of moving IoU and our network runs at 50 fps which is very crucial for autonomous driving. The moving object segmentation problem needed to be studied further, to explore better fusion strategies to improve the accuracy and efficiency and there is every need to build large varied datasets for moving object segmentation.

[1] https://youtu.be/S56btzltRTM

References

[1] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019.

[2] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[3] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[4] Hazem Rashed, Mohamed Ramzy, Victor Vaquero, Ahmad El Sallab, Ganesh Sistu, and Senthil Yogamani. Fusemodnet: Real-time camera and lidar based moving object detection for robust low-light autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[5] Mennatullah Siam, Heba Mahgoub, Mohamed Zahran, Senthil Yogamani, Martin Jagersand, and Ahmad El-Sallab. Modnet: Motion and appearance based moving object detection network for autonomous driving. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2859–2864. IEEE, 2018.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

CT Radiomic Features for a Prostate Cancer Evaluation Framework

Bruno Mendes
brunomendes81@gmail.com
Inês Domingues
inesdomingues@gmail.com
João Santos
joao.santos@ipoporto.min-saude.pt

Faculdade de Engenharia da Universidade do Porto
Portugal
Medical Physics, Radiobiology and Radiation Protection Group, IPO Porto Research Centre (CI-IPOP), Portugal
Instituto de Ciências Biomédicas Abel Salazar
Porto, Portugal

Abstract

Transrectal Ultrasound Guided Biopsy (TRUS) is the principal method to diagnose Prostate Cancer (PCa) following a histological examination by observing cell pattern irregularities and assigning the Gleason Score (GS) according to the recommended guidelines. This procedure presents sampling errors and, being invasive may cause complications to the patients. In this work, we evaluated the use of data-characterization algorithms (radiomics) from Computed Tomography (CT) images for PCa aggressiveness assessment. The extracted features show a poor correlation with the GS. But, when using the Principal Component Analysis (PCA), results improve dramatically and show great promise.

1 Introduction

The first described PCa case goes back to 1853, when John Adams, a surgeon at the London Hospital, followed a histological examination for cirrhosis of the prostate gland. He reported the condition as an orphan disease. In 2020, it was the second most frequent malignancy, with 1.414.259 new cases and responsible for 7.3% of all cancer deaths in men [7].

PCa is usually asymptomatic at an early stage and screened by Digital Rectal Examination (DRE) and Prostate Specific Antigen (PSA) blood test. The principal method to diagnose PCa is the TRUS with samples taken mainly from the peripheral zone [2]. The pathologist identifies the two most predominant sets of patterns. He then assigns a score of one if prostate cells are uniformly packed, up to a grade of 5 depending on pattern irregularity. The sum of both is designated the GS and is proportional to PCa aggressiveness. Several studies showed that a GS of $7 = 4+3$ has a worse prognosis than a GS of $7 = 3+4$. Taking this into account, Epstein et al. [3] proposed a new stratification by Grade Group (GG), as shown in Figure 1. This new grading system provides the potential to reduce the overtreatment of indolent cancer and reflects the high heterogeneity of PCa [3].

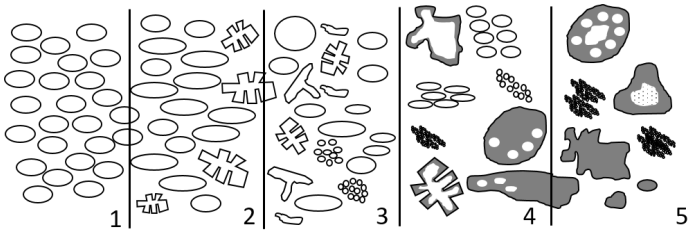


Figure 1: Stratification by GG. Adapted from [3].

Heterogeneous solid cancers may limit invasive biopsies but open an opportunity to medical imaging. Particularly when significant differences in protein expression patterns proved to correlate to radiographic findings [5]. CT images have a high spatial resolution allowing the evaluation of density, shape and texture characteristics. Radiomics, the extraction of features from radiographic images using data-characterization algorithms, may provide a valuable tool for PCa grading during External Beam Radiotherapy Treatment (EBRT). The hypothesis behind radiomics is that quantitative analysis of medical images may provide a similar prognosis power as phenotypes and gene protein signatures. The present study aims to evaluate the potential use of radiomics to predict PCa aggressiveness using CT images.

2 Materials and Methods

2.1 The image dataset

This preliminary retrospective research used treatment plans available at Instituto Português de Oncologia do Porto Francisco Gentil (IPO-PORTO). All patients had undergone a CT scan as part of the EBRT treatment. The GS was also available as part of the initial grading process. The image dataset had CT images from 44 patients following a 3-fold GS risk group stratification, as suggested by Epstein et al. [3] and presented in Table 1. Experts at the institution had delineated all Volume Of Interest (VOI) and Organ At Risk (OAR). The Clinical Target Volume (CTV) was chosen as the feature extraction region because it presents the most clinical and pathological information.

Each study, CT series and the CTV, was visualized in the 3D Slicer [4] platform and converted to Nearly raw raster data (Nrrd) volume. This format only stores pixel information and thus assures proper anonymization. The final dataset consisted of 44 volumes and corresponding CTVs in the Nrrd format. The number of patients per Risk Group (RG) shows a very imbalanced dataset even following a slice by slice approach, as shown in Table 1. We attempted a slice by slice approach for comparison with the more conventional and recommended volume approach. The goal was to capture minor irregularities and variations in the prostate gland.

Table 1: Number of cases and images per risk group.

RG	GG	GS	# Cases	# Images
Low/Very Low	1	≤6	3	56
Intermediate (Favorable/Unfavorable)	2	7 (3+4)	31	664
	3	7 (4+3)		
High/Very High	4	8	10	209
	5	9-10		
Total			**44**	**929**

2.2 Feature Extraction and Selection

We used the python library pyradiomics [8], a highly tested and maintained open-source platform, to extract features from the CTV. Most pyradiomics features comply with the Image Biomarker Standardisation Initiative (IBSI), an independent international collaboration that aims at standardizing the extraction of image biomarkers for high-throughput quantitative image analysis (radiomics).

Features were standardized by removing the mean and scaling to unit variance. Each image or volume descriptor represents a point in the feature space. But some are highly correlated, which means overlapped axis. To overcome this issue, we used PCA, which projects the data points to an uncorrelated and orthogonal axis to maximize variance [6]. Dimensionality reduction occurs with the selection of higher variance components.

The dataset is quite imbalanced, especially for the minority class (Low/Very Low). To overcome this issue, we generated synthetic features applying the Synthetic Minority Oversampling Technique (SMOTE) [1].

2.3 Model Building and Classification

The adopted methodology allows having a dataset with multiple image features labelled with a particular output, the GS. CT images are not the *de facto standard* for PCa evaluation, so we attempted a more conservative approach. We used an One-vs-rest (OvR) multiclass strategy with Support Vector Machine (SVM) as a baseline. With this approach,

we fitted one classifier per class against all the others. To assess performance, we computed the Area Under the Receiver Operating Characteristic (AUROC) curve up to a maximum number of components given by $min(\#samples \times 0.8, \#features)$, and the Average Precision (AP) for 30 runs.

To built the slice by slice model, we considered a two splits group k-fold, assuring that the testing and training sets do not have repeated cases. We generated synthetic data on the training set only, and the test size was 20% of all slices or volumes. The feature extraction and model building pipeline encompasses several steps, as exemplified in Figure 2.

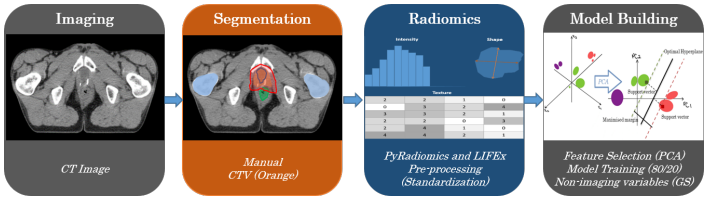


Figure 2: Radiomics Pipeline.

3 Results

This work followed an image and volume feature extraction. Shape-based (3D) features are not extracted on a slice by slice approach. Others suffer dimension constraints for a more accurate calculation. For example, the second-order joint probability function of an image region, defined as Gray Level Co-occurrence Matrix (GLCM), uses a maximum of 49 unique angles to perform calculations when in 3D [8]. Although PyRadiomics computed the features on a slice basis by reducing the third dimension to one, the obtained values lack the depth required for a more accurate estimation.

The fundamental ideal behind radiomics is to find a signature. In other words, a feature or a set of features that show a high correlation to the GS. Unfortunately, with CT images, we were not able to find such a signature. The extracted features show a poor correlation with the GS and a strong inter-correlation, as shown in Figure 3.

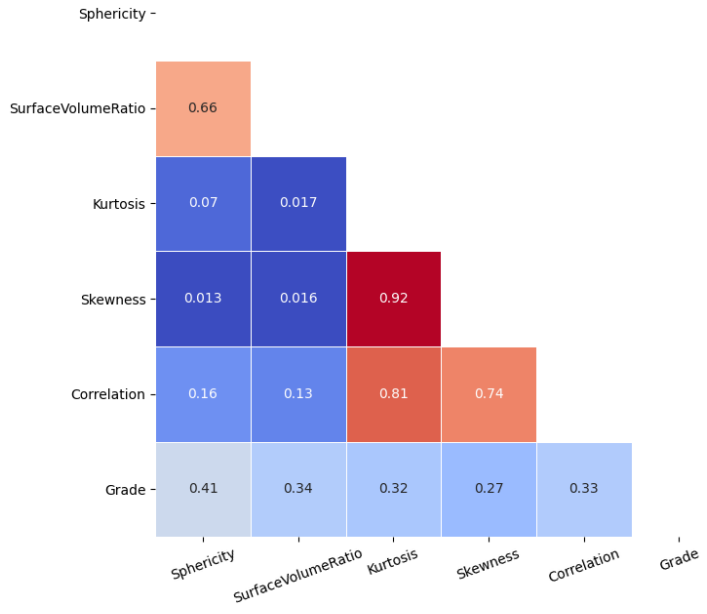


Figure 3: Top 5 most GS correlated features.

To overcome this, we used the PCA. To evaluate the parameters that provided the best output, we computed the AUROC for each available SVM kernel and the number of principal components. We obtained a maximum mean AUROC value using 75 and 6 principal components for the slice and volume approach respectively. The linear kernel provided the best results for both methods. Adding volumetric information did provide better output, as shown in Table 2. The full volume approach did provide better overall performance for all classes with fewer components mostly since some features provide more accurate values with the addition of angles and neighbours on a 3D basis.

Table 2: Best AUROC values.

	Low/Very Low	Intermediate	High/Very High	AP
Per slice	0.77	0.63	0.75	0.78
Volume	0.88	0.79	0.88	0.85

4 Conclusions

The low soft-tissue contrast and the lack of metabolic manifestation of CT provides a challenge for a possible radiomic signature. With the PCA, we were able to reduce dimensionality by linearly combining features with higher variance. Also, we achieved a better sparsity. As a downside, we lose the ability to identify the radiomic signature for PCa aggressiveness.

The dataset is small and imbalanced as shown in Table 1 with the "Low/Very Low" class being under-represented. This issue was addressed in the slice by slice approach by generating synthetic data with SMOTE but still a bigger dataset is needed for a full volume approach. Several features are yet to be considered, such as the ones obtained from the derived images after applying Laplacian of Gaussian (LoG) and wavelet filters.

PCa grading is a complex task with multiple variables to be evaluated. The present study may provide the baseline to develop an accurate classifier to predict PCa aggressiveness during treatment using CT images. Such a tool may improve decision outcomes and avoid overdiagnosis and overtreatment.

References

[1] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[2] Jean-Luc Descotes. Diagnosis of prostate cancer. *Asian journal of urology*, 6(2):129–136, 2019.

[3] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, and Peter A Humphrey. The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American journal of surgical pathology*, 40(2):244–252, 2016.

[4] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3D Slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9):1323–1341, 2012.

[5] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.

[6] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of significance: Principal component analysis, 2017.

[7] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: GLOBOCAN Estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 2021.

[8] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

Detection of *EGFR*-related patterns in lung cancer CT images: a holistic approach

Francisco Silva[12]
francisco.c.silva@inesctec.pt
Tania Pereira[1]
tania.pereira@inesctec.pt
António Cunha[13]
antonio.cunha@inesctec.pt
Hélder P. Oliveira[12]
helder.p.oliveira@inesctec.pt

[1] Faculty of Sciences of University Porto, Porto, Portugal
[2] INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal
[3] UTAD - University of Trás-os-Montes and Alto Douro, Vila Real, Portugal

Abstract

The late diagnosis is one of the main factors responsible for the high death rate related to lung cancer. To increase survival chances, advances in individualized treatments based on genetic profiling have shown to provide better control on cancer response. To assess the tumor characterisation, medical imaging data offers valuable information on the tumor, opening opportunities to explore genotype-related imaging patterns in a non-invasive way. This work aims to study the relevance of physiological features captured from computed tomography images, using three regions of interest (ROI) to predict the Epidermal growth factor receptor (*EGFR*) mutation status. Results showed that extending the analysis beyond the nodule allowed to capture more relevant *EGFR*-related patterns, with best performance achieved using the lung with nodule ROI. This comparative study contributes to the discussion about how extensive the imaging patterns associated with cancer development are, and their importance for more accurate AI-based solutions.

1 Introduction

Lung cancer leads the cancer-related mortality rate numbers [5]. Considering the high number of patients diagnosed with advanced-stage disease, the selection of a personalized treatment plan based on genetic profile of the patient has shown the potential to increase survival rates [3]. By acting on specific molecular targets responsible for growth and cancer progression, these therapies allow to increase the success of the treatment response. However, the development of these therapies requires the identification of specific genetic biomarkers, being the Epidermal growth factor receptor (*EGFR*) the most well-studied oncogene in lung cancer [6]. Traditionally, this identification is performed by molecular tests using the tissues extracted during the biopsy. Recently, less invasive and more automatic techniques based on computed tomography (CT) analysis have been developed, which decreases the risk for the patients.

The lack of publicly available data has been the biggest limitation in the development of computer-aided systems for lung cancer characterization. However, several works have already suggested the possibility of finding imaging patterns in CT data correlated with the *EGFR* mutation staus. The use of qualitative features annotated by radiologists showed to enable an accurate assessment of *EGFR* [2, 4], by including features from multiple lung structures, not focusing only on the nodule region. Deep learning approaches have also been proposed for this task, including methodologies based on transfer learning [9], ensemble strategies [10] and 3D CNN models trained from scratch [11]. Although these works analysed different regions for *EGFR* prediction, the focus has been kept centered on the nodule. To allow a more comprehensive analysis of the lung pathological processes associated with cancer development, it might be necessary to explore larger regions of interest, not restricting the studied data to the nodule region.

Considering this, two main *EGFR*-related challenges represent the focus of this work: the study of the ROI that allows to capture more relevant imaging patterns associated with *EGFR* mutation status, aiming to achieve a more comprehensive analysis of the biological problem, and the development of an approach to overcome the lack of massive annotated databases. The current paper is an adaptation of our previously published work [1].

2 Materials and Methods

2.1 Datasets

The LIDC-IDRI [7] is a lung cancer screening dataset that comprises thoracic CT scans for a total of 1010 patients, alongside with annotated nodules belonging to one of three classes: a) nodule $\geq$ 3 mm; b) nodule $<$ 3 mm or c) non-nodule $\geq$ 3 mm. A total of 2669 lesions were classified as larger than 3 mm by at least one clinician, for which was made available nodule contours marked by each radiologist.

The NSCLC-Radiogenomics dataset [8] is a public available database with CT images for a cohort of patients with non-small cell lung cancer, being the only public dataset comprising paired information on lung cancer-related gene mutation status and CT data. Only 116 patients were included due to a required valid *EGFR* mutational test result and the availability of tumor binary mask. From these, 23 (20%) belonged to the *Mutant* class, and 93 (80%) to the *Wildtype* class.

2.2 Data Preparation

To standardize image representations, the pixel spacing was set to [1.00, 1.00, 1.00] mm, computing each CT dimensions to match the new spacing values. Additionally, each pixel intensity value, measured in the Hounsfield Units (HU) scale, was normalized using the *min-max* algorith, and values under -1000 HU were transformed into 0 and values above 400 HU were transformed into 1, mapping all values in the middle into the [0, 1] range (see examples of processed input data in Fig.1 below).

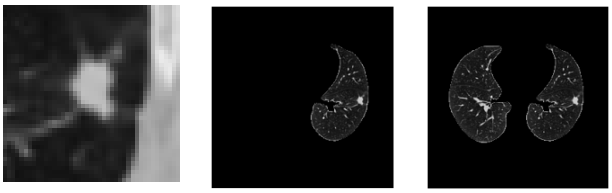


Figure 1: Example of each considered ROI: nodule, lung with nodule and both lungs, extracted from an NSCLC-Radiogenomics [8] patient.

2.3 Methodology

Being widely explored to overcome the lack of public data in the medical imaging field, transfer learning techniques allow the use of more complex architectures by using networks pre-trained on massive datasets. Given the scarce dataset size available for *EGFR* prediction, we considered an alternative approach that consists of developing and training the a convolutional autoencoder (CAE) in the same domain of the final task. This transfer learning strategy was chosen based on the intuition that the trained encoder would be capable of achieving the necessary general knowledge in the lung cancer domain, and intends to explore the relevance of the learned patterns while reconstructing input images.

To the classification block, fully-connected layers were stacked on top of the pre-trained encoder. Given the ability of this neural network-based classifier to backpropagate the prediction error to the encoder layers, it was possible to fine-tune the lower-level layers of the encoder, which helped the model to learn to detect the most useful *EGFR*-related patterns (see Figure 2).

Considering the experiments conducted, both the nodule and lung analysis consisted of a feature learning task for the CAE training, and a

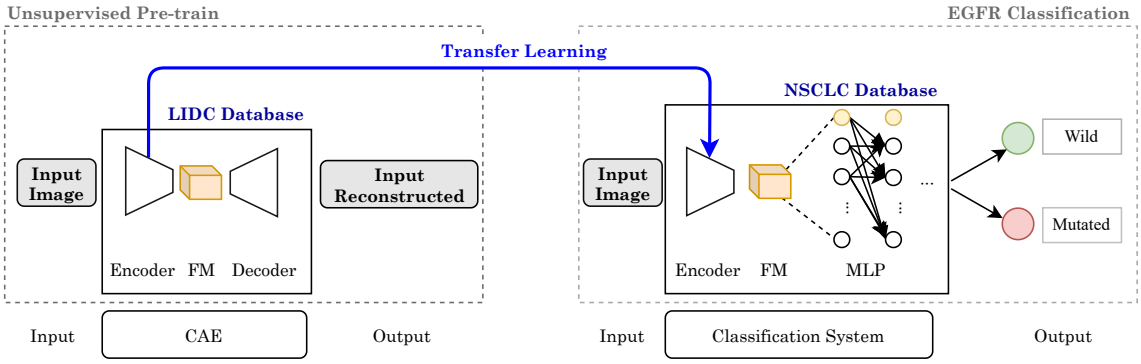


Figure 2: Methodology overview: the unsupervised pre-training of the CAE to be used as feature extractor, and an end-to-end classifier to predict the *EGFR* mutation status. Transfer learning allows to reuse the encoder for the extraction of feature maps (FM) to predict the *EGFR* mutation status.

classification task, where by making advantage of transfer learning techniques, the *EGFR* mutation status was predicted on the correspondent ROI: nodule, lung containing nodule and both lungs.

3 Results and Discussion

Table 1 summarizes the achieved results, computed over 20 random train-test splits for each proposed experiment.

Experiment (ROI)	AUC
Nodule	0.51 ± 0.06
Lung containing nodule	0.68 ± 0.08
Both lungs	0.60 ± 0.10

Table 1: Classification results for *EGFR* mutation status prediction. AUC values are depicted as mean $\pm$ standard deviation over 20 random train-test splits for each experiment.

Analysing the results achieved in each experiment, the worst classifier was developed when analysing only the nodule region, not being capable to assess the *EGFR* mutation status (AUC = 0.51). When extending the ROI to the entire lung axial section, the best classification performance was achieved when including the lung that contained the nodule (AUC = 0.68). The achieved results showed the relevance of the learned features when trying to reconstruct the input image. To the best of gathered knowledge, no other deep learning-based work attempted to assess the *EGFR* mutation status using a holistic analysis.

The achieved results indicate a direction for future works dedicated to lung cancer characterization. A direct comparison with related studies, in particular, the ones with higher ability to predict the *EGFR* mutation status would not bring a fair discussion point to this investigation, given the fact that the proposed approach addresses the lack of publicly available data to perform this task, which is not a factor that constrained the contribution of those studies. Nevertheless, the current work contributes to the discussion about how complex and extensive are the imaging patterns associated with cancer development. Only few recent works showed the relevance of using information from other parts of the lung to predict the *EGFR* mutation status associated with lung cancer [2, 4], pointing out the importance to use comprehensive approaches that take into consideration more elements to characterize these extremely complex physiological processes. However, to develop models capable to make a more comprehensive analysis with further potentially relevant information, more representative data of the population affected by lung cancer is needed to enable such abstract and complex patterns detection.

4 Conclusions

This study proposed an approach based on unsupervised transfer learning techniques for the classification of *EGFR* mutation status. Different regions of interest were analysed in order to study the relevance of information from all the lung structures in this complex classification task. The obtained results showed that information from more extensive regions on the lung containing the nodule allow to detect imaging patterns that might be relevant for lung cancer characterization.

Acknowledgment

We acknowledged the National Cancer Institute and the Foundation for the National Institutes of Health for the free publicly available LIDC-IDRI Database used in this work. We acknowledged The Cancer Imaging Archive (TCIA) for the open-access NSCLC-Radiogenomics dataset publicly available.

Funding

This work was partially funded by the Project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership and by the PhD FCT scholarship 2021.05767.BD.

References

[1] Francisco Silva *et al.* EGFR Assessment in Lung Cancer CT Images: Analysis of Local and Holistic Regions of Interest Using Deep Unsupervised Transfer Learning. *IEEE Access*, 2021.

[2] Gil Pinheiro *et al.* Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. *Scientific Reports*, 2020.

[3] Min Yuan *et al.* The emerging treatment landscape of targeted therapy in non-small-cell lung cancer. *Signal Transduction and Targeted Therapy*, 2019.

[4] Olivier Gevaert *et al.* Predictive radiogenomics modeling of EGFR mutation status in lung cancer. Technical report, 2017.

[5] Rebecca L. Siegel *et al.* Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 2017.

[6] S. Jorge *et al.* Epidermal growth factor receptor (EGFR) mutations in lung cancer: Preclinical and clinical data. *Brazilian J. Med. Biol. Res.*, 47(11):929–939, 2014.

[7] Samuel G. Armato *et al.* The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med. Phys.*, 38(2):915–931, 2011.

[8] Shaimaa Bakr *et al.* Data descriptor: A radiogenomic dataset of non-small cell lung cancer. *Sci. Data*, 5, 2018.

[9] Shuo Wang *et al.* Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. *Eur. Respir. J.*, 53(3), 2019.

[10] Silvia Moreno *et al.* A Radiogenomics Ensemble to Predict EGFR and KRAS Mutations in NSCLC. *Tomogr. (Ann Arbor, Mich.)*, 7(2): 154–168, 2021.

[11] Wei Zhao *et al.* Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3D deep learning. *Cancer Med.*, 8(7):3532–3543, 2019.

Machine learning approach for perfusion assessment of synthetic myocardial SPECT images

Sérgio Figueiredo[1,2]
sergio.r.figueiredo@tecnico.ulisboa.pt

Ana L. N. Fred[3]
afred@lx.it.pt

J. Miguel Sanches[1]
jmrs@tecnico.ulisboa.pt

[1]Institute for Systems and Robotics (ISR), LARSyS, Instituto Superior Técnico, Department of Bioengineering, Universidade de Lisboa

[2]H&TRC – Health & Technology Research Center, ESTeSL – Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa

[3]Instituto de Telecomunicações Instituto Superior Técnico, Department of Bioengineering, Lisbon

Abstract

Coronary artery disease (CAD) is the leading cause of premature mortality. SPECT myocardial perfusion imaging (MPI) is the cardinal modality for the diagnosis of significant CAD but is difficult to interpret mostly due to the collimator spatial blurring. This project aims to develop a classifier that can determine if a SA SPECT MPI has normal or abnormal perfusion. For this, a dataset that tries to mimic real SA SPECT images was synthetised and representative features were extracted. The Random Forest (RF), Logistic Regression and K-nearest Neighbors classifiers were used. RF achieved the better performance (Accuracy = 96,82%), supporting that these techniques might contribute to the assessment of SPECT MPI.

1 Introduction

Cardiovascular diseases (CVD), particularly coronary artery disease (CAD), remain the world leading cause of premature mortality [1]. Single-photon emission computed tomography (SPECT) uses gamma emitter radiopharmaceuticals for the study of myocardial perfusion and continues to be the most cardinal non-invasive modality for diagnosis and risk stratification of hemodynamically significant CAD [2].

Typically, the acquired SPECT raw-data is reconstructed using filtered-back projection (FBP) or iterative algorithms (*e.g.*, OSEM-MLEM), sectioning the information into vertical long-axis (VLA), horizontal long-axis (HLA) and short-axis (SA) planes, based on the left ventricular (LV) cardiac planes (Figure 1a and 1b) [3]. The interpretation of these images is commonly performed by expert readers through software, using the SA slices in a 17 segment polar map of the LV (Figure 1c) to apply perfusion quantification [3] and detect abnormal areas, categorizing them as a function of defect extent [4] [5].

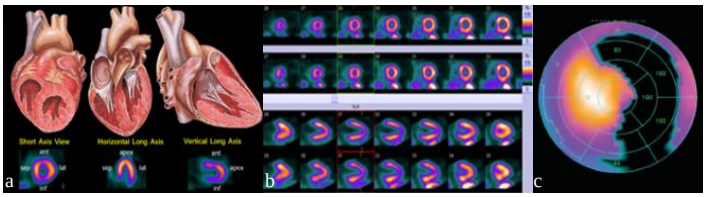


Figure 1. Typical mid-SA, VLA and HLA cardiac post-reconstruction planes **(a)** and SA and VLA planes (**b**). LV polar map perfusion defect quantification of a stress ^{99m}Tc-MIBI scan **(c)**.

In practice, the initial assessment of the SA, VLA and HLA slices is visually carried out by nuclear medicine/cardiologist expert, and additionally correlated to the perfusion quantitative parameters [4] [6]. Consequently, interpreting these blurred images is challenging and is a time-consuming task, naturally linked to significant inter and intra-observer variability [6]. Nevertheless, transforming knowledge of expert readers to appropriate image processing methods to automatic classify SPECT MPI images, is still limited reported [6], particularly using machine learning (ML) algorithms.

2 Problem Formulation and Motivation

Cardiac SPECT MPI is difficult to interpret due to certain degrading factors, principally related to collimator depth-dependent spatial blurring, attenuation, scatter and low counts, mainly associated to high levels of noise and poor contrast [7], characterized by Poisson noise and Gaussian blur [4].

The image formation model can be approximated by the following convolution operation:

$$I_B(x,y,z) = O(x,y,z) * h(x,y,z) \quad (1)$$

where * denotes the convolution operator, $h(x,y,z)$ is the three-dimensional point spread function (PSF) of the acquisition system, I_B is the blurred observation and O is the unknown radioactive activity to be estimated.

Considering this assumption, this project aims to classify SA SPECT myocardial perfusion synthetic images as normal or "abnormal" perfusion, in order to improve the accuracy of the diagnosis, in comparison to the conventional visual inspection approach.

3 Methods and Materials

The dataset (DS) contains a total of 700 simulated SA SPECT MPI synthetic images: 350 normal perfusion (N) and 350 abnormal perfusion (ABN), distorted with realistic blur and corrupted by Poisson noise, as analogous to the mid-SAs SPECT slices [8].

To mimic the gaussian blur, the standard deviation (σ) value was estimated using mid-SA slice SPECT MPI from a real SPECT MPI study (demonstration patient available from Siemens Healthineers, School of Health Technology (ESTeSL), Politécnico de Lisboa). A cross-sectional profile of the ROI (Region of Interest) was computed on the septal wall (Figure 2a), to generate a revolution gaussian kernel with the parameter σ = 12 and size 41 (Figure 2b).

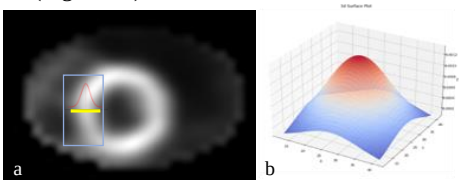


Figure 2. Cross-sectional profile of the ROI (yellow) on the septal wall (blue) **(a)** and 3D Gaussian function (σ = 12) **(b)** obtained from **(a)**.

The 8-bit 256x256 images of the DS represent the perfusion of typical LV myocardial 8 to 15 mm thickness, assuming a pixel value range between 200 and 255 for N (Figure 3a and 3b), and a range between 15 and 130 for the ABN (Figure 3c and 3d). A random failure range angle was created to simulate the arc length associated to the deficit and a blending linear operator was used in order to simulate the perfusion defect as a non-zero pixel value, similar to real SA SPECT ((Figure 3c and 3d). Here we took two images, the first intends to represent the normal perfusion and is given a weight of 0.7 (α) and the second image is given 0.3 (β), representing the perfusion defect. These parameters control the transition between one image to another and applies the following equation to the final image:

$$g(x,y) = (1-\alpha)f_0(x,y) + \alpha f_1(x,y) \quad (2)$$

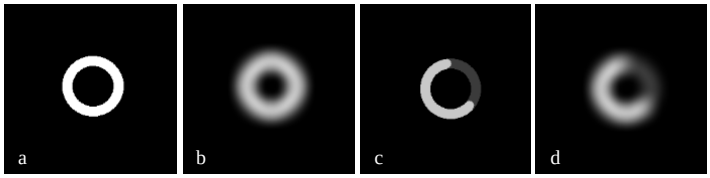


Figure 3. LV SA SPECT MPI synthetic images. **(a)** Normal perfusion ground truth. **(b)** Blurred normal perfusion. **(c)** Gradient perfusion defect ground truth. **(d)** Blurred gradient perfusion defect.

3.1. Segmentation

To extract the features for classification, different regions of the images were segmented, generating two masks. To obtain the object mask (Omk), *i.e.*, to isolate the whole mid-myocardial SA SPECT pixels, the Niblack and Sauvola method [9] was applied, since they are local thresholding techniques that are useful for images where the background is not uniform.

To obtain the normal mask (Nmk), *i.e.*, to isolate the normal perfusion region, an adaptive thresholding technique based on Otsu's thresholding algorithm was used with acceptable results.

*This work was supported by Portuguese funds through FCT (Fundação para a Ciência e Tecnologia) through the projects reference UIDP/50009/2020 and through the reference UID/EEA/50009/2019, LARSyS - FCT Plurianual funding 2020-2023.

3.2. Classification

Considering the characteristics of the SA SPECT images, the first computed feature was the pixel values intensities (PV) (Table 1). Since the perfusion deficits assume diverse extension and might be located at different quadrants of the SA plane, the total number of pixels of the normal perfusion region (TNP) and Asymmetry (AS), particularly, the Symmetry Left-Right (SLR) and Symmetry Top-Bottom (STB) were computed as features (Table 1). After the feature extraction, three models were implemented, to discriminate the two classes, *i.e.*, normal and abnormal perfusion. A supervised learning approach was used by making use of some classical classification models, particularly, Logistic Regression (LR), Random Forest (RF) and K-Nearest Neighbours (KNN), as they are commonly used ML techniques in the field of cardiac imaging and diagnosis [10].

Table 1. Features extracted and correspondently description.

Feature		Description
PV		Is related to the mean value of the pixel's intensities of the whole object in the image.
TNP		Reflects the total number of pixels of the normal perfusion region.
AS	**SLR**	The asymmetry (AS) of the image reflects the existence of an abnormal perfusion region. Each image was divided in two, horizontally (STB) and vertically (SLR), and one of the sides is inverted and overlapped in the other. The ratio between the intersection and the union of the halves gives a coefficient, higher as the level of symmetry.
	STB	

The default *scikit-learn* stratified *k*-fold cross-validator was used, *i.e.*, 10 and 3 cross-validation folds were correspondingly assigned to LR and KNN models, and RF model. The dataset was split into train and test sets, with percentages of 45-55%, respectively, guarantying a training with a balanced dataset.

To estimate the best set of corresponding hyperparameters for the LR, RF and KNN models, the Grid search (GS) and the Randomized search (RS) methods were applied, since they are one of the most commonly used tools to hyper-parameter space tunning [11]. After that, the optimal hyperparameters were used to retrain each model with the whole DS.

The performance metrics accuracy (acc), sensitivity (sen), specificity (spec), F1-Score and ROC AUC were used to assess the performance of the LR, RF and KNN models.

4 Results and Discussion

In this work, three different classification models, namely, LR, RF and KNN, were introduced to identify perfusion abnormalities using a synthetic SA SPECT MPI dataset.

A comparison of the test classification performance related to the LR, RF and KNN models is registered in Table 2.

Table 2. Classification performance (%) for the LR, RF and KNN models.

Classifier	Accuracy	Sensitivity	Specificity	F1 Score	ROC AUC
LR	95,87	**95,40**	96,45	96,23	99,38
RF	**96,82**	94,25	**100,0**	**97,04**	**99,94**
KNN	88,88	86,78	91,49	89,61	96,75

In Table 2 it is noticeable that all models provide good results, predominantly assuming values superior to 80%, overall. However, related to the Acc, it can be observed that the RF model obtained the highest value (96,82%) and the KNN the lowest one (88,88%), meaning that the RF model is the model that shows a more robust performance.

Related to sensitivity, the LR model achieves the highest value (95,40%) and KNN still remains with inferior results (86,78%), while RF obtained a value of 94,25%. Concerning the specificity metric, the RF models achieves a value of 100%, while the lowest result is associated to the KNN (91,49%), and LR obtained a value of 96,45%.

In this sense, considering that a higher Sen implies a higher rate of true negatives and higher Spec implies higher TPR, it is perceptible that the RF model is more specific since the number of false positives obtained in the confusion matrix is 0, as the true positive is 164. In case of detection of perfusion deficit, this implies that the RF model can identify a high number of true positives, ensuring that none is identifying as having a false perfusion deficit. Comparable, related to the LR model, the confusion matrix demonstrates a value of 166 and 5, for the true positive and false positive, respectively. Nevertheless, related to this kind of problem, this trained LR model can lead to unnecessary invasive angiographies [6], as the patient is classified as positive while truly is false (false positive).

Regarding the F1-Score and AUC ROC metrics, it is observable that all models achieve very good results, revealing a good inter-class separability, *i.e.*, distinguishing between N and ABN perfusion.

The LR model is a simple and explainable classifier that achieved excellent results (Table 2), and might be a good baseline related to the other remarkably complex models. Further, the KNN models revealed inferior results but acceptable.

Generally, the RF model provided the best performance results and might be the most suitable ML approach to classify the SA SPECT MPI images. This is supported by the work of Ricciardi *et al.* [12] using 10,265 patients with suspected or known CAD, undergoing SPECT MPI, demonstrated that the Random Forests obtained the highest accuracy (> 95%), while KNN the lowest recall and sensitivity (79.2%).

5 Conclusions

The Random Forest model provided the best performance to classify the SA SPECT MPI synthetic images, accomplishing the higher and excellent results, when compared to the LR and KNN models.

The promising results of this study support the notion that this ML approach can be used to automatic quantify the myocardial perfusion and might contribute to the interpretation of SPECT MPI. Nevertheless, further research using realistic SA SPECT MPI data is required to validate this scenario, simulating the typical the phenomenon of sub-diaphragmatic activity area (*e.g.*, liver or intestine) that may distort the segmentation of the myocardium LV inferior wall [4].

References

[1] G. A. Roth *et al.*, "Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study," *J. Am. Coll. Cardiol.*, vol. 76, no. 25, pp. 2982–3021, 2020, doi: 10.1016/j.jacc.2020.11.010.

[2] D. D. Watson and D. K. Glover, "Chapter 1 - Overview of Tracer Kinetics and Cellular Mechanisms of Uptake," in *Clinical Nuclear Cardiology (Fourth Edition)*, Fourth Edi., B. L. Zaret and G. A. Beller, Eds. Philadelphia: Mosby, 2010, pp. 3–13.

[3] T. L. Faber, J. I. Chen, and E. V Garcia, "Chapter 4 - SPECT Processing, Quantification, and Display," in *Clinical Nuclear Cardiology (Fourth Edition)*, Fourth Edi., B. L. Zaret and G. A. Beller, Eds. Philadelphia: Mosby, 2010, pp. 53–71.

[4] S. Dorbala *et al.*, "Single Photon Emission Computed Tomography (SPECT) Myocardial Perfusion Imaging Guidelines: Instrumentation, Acquisition, Processing, and Interpretation," *J. Nucl. Cardiol.*, vol. 25, no. 5, pp. 1784–1846, 2018, doi: 10.1007/s12350-018-1283-y.

[5] P. J. Slomka, D. S. Berman, and G. Germano, "Quantification of Myocardial Perfusion," in *Clinical Gated Cardiac SPECT*, John Wiley & Sons, Ltd, 2006, pp. 69–91.

[6] S. Kaplan Berkaya, I. Ak Sivrikoz, and S. Gunal, "Classification models for SPECT myocardial perfusion imaging," *Comput. Biol. Med.*, vol. 123, no. June, p. 103893, 2020, doi: 10.1016/j.compbiomed.2020.103893.

[7] G. Germano and D. S. Berman, "Physics and Technical Aspects of Gated Myocardial Perfusion SPECT," in *Clinical Gated Cardiac SPECT*, John Wiley & Sons, Ltd, 2006, pp. 27–45.

[8] O. S. Hanafy, M. M. Khalil, I. M. Khater, and H. S. Mohammed, "Development of a new Python-based cardiac phantom for myocardial SPECT imaging," *Ann. Nucl. Med.*, vol. 35, no. 1, pp. 47–58, 2021, doi: 10.1007/s12149-020-01534-y.

[9] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognit.*, vol. 33, no. 2, pp. 225–236, 2000, doi: 10.1016/S0031-3203(99)00055-2.

[10] C. Martin-Isla *et al.*, "Image-Based Cardiac Diagnosis With Machine Learning: A Review," *Front. Cardiovasc. Med.*, vol. 7, no. January, pp. 1–19, 2020, doi: 10.3389/fcvm.2020.00001.

[11] L. Yang, A. Shami, T. Yu, and H. Zhu, "Hyper-Parameter Optimization : A Review of Algorithms," no. July, pp. 1–56, 2020.

[12] C. Ricciardi *et al.*, "Application of data mining in a cohort of Italian subjects undergoing myocardial perfusion imaging at an academic medical center," *Comput. Methods Programs Biomed.*, vol. 189, p. 105343, 2020, doi: 10.1016/j.cmpb.2020.105343.

Comparison of bladder segmentation techniques in CT scans

Ana Couto
AnaSofia1995@hotmail.com
Inês Domingues
inesdomingues@gmail.com
João Santos
joao.santos@ipoporto.min-saude.pt

Faculdade de Ciências da Universidade do Porto
Portugal
Medical Physics, Radiobiology and Radiation Protection Group, IPO Porto Research Centre (CI-IPOP), Portugal
Instituto de Ciência Biomédicas Abel Salazar
Porto, Portugal

Abstract

Radiotherapy takes a very important role in cancer treatment. One of its necessary steps is to segment the Organs at Risk. This process is currently done manually, which is time consuming and subject to human error. With the goal of helping the specialists and improving segmentation accuracy, some algorithms were tested to segment the bladder in 47 Computerized Tomography scans from patients with prostate cancer provided by the Institute of Oncology of Porto. The four algorithms were evaluated and the Dice obtained for applying Clustering, U-Net, Active Contours and Graph Based were 24%, 6%, 26% and 29%, respectively, in the data set with the mask based on HU values. For anatomic mask, the same metric for the same algorithms were 22%, 80%, 31% and 20%, respectively.

1 Introduction

Surgery and radiotherapy are the most efficient and used treatments for cancer. 60% of the patients submitted to radiotherapy are treated with curative intent but it also has an important role in the reduction of symptoms [1]. In order to find a balance between eradicating the tumour and sparing the surrounding tissues, it is important to observe in detail where the tumour ends and the surrounding organs begin. This procedure is made manually, which is time consuming and subject to human error, namely variability between different contours made by different specialists (inter-variability) and variability between different contours made by the same specialist (intra-variability).

The main goal of this work is to study different segmentation algorithms to apply to prostate cancer patients' CT scans provided by the Institute of Oncology of Porto (IPO). The organ of interest is the bladder. By doing so, it is intended to reduce the time consuming manual segmentation and improve its accuracy.

The remaining of this paper is organised as follows. Section 2 describes the segmentation techniques evaluated and compared in the present work. The dataset is presented in Section 3. Section 4 gives the results while in Section 5 some conclusions and directions for future work are given.

2 Segmentation Methods

Two pre-procedures were developed in order to detect the Region of Interest (RoI): one based on thresholding each CT using Hounsfield Unit (HU) value of the bladder and another one based on the anatomy. These pre-procedures were applied to each of the 71 bladders. The segmentation algorithms are described next.

2.1 Clustering

The clustering algorithm used was K-means. k values between 5 and 195 were tested by steps of 10. The next step is to automatically choose one cluster. With this purpose, 5 features were extracted from each cluster: volume, diameter and the maximum, minimum and mean intensity. Several classifiers were trained: Classification Trees (CT), Discriminant Analysis (DA), k Nearest Neighbors (kNN), Naive Bayes (NB), Support Vector Machine (SVM), Classification Ensembles (CE), Classification Tree Ensembles (CTE) and RUSboost. Two different post processing were applied to the selected cluster. In the first one, the regions and holes were filled by implementing flood fill. The second approach consists of choosing the biggest connect region. The full pipeline of the Clustering segmentation is shown in Figure 1.

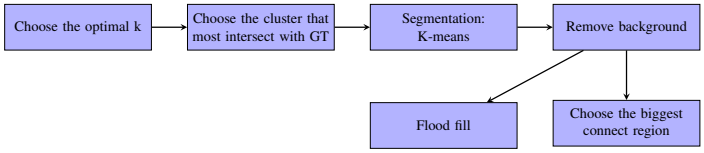


Figure 1: Clustering Process

2.2 U-Net

This algorithm was based and adapted from [2]. Training a network on the full input volume is impractical due to the amount of memory needed to store and process 3-D volumes. This problem is solved by training the network on image patches extracted from the ground truth images. To prevent overfitting due to data limited size, the training and validation data were augmented by randomly rotating and reflecting training data to make the training more robust. In order to avoid border artefacts when using the overlap-tile strategy for prediction of the test volumes, valid convolution padding was specified. The overlap-tile strategy was used to predict the labels for each test volume. The full pipeline of the U-Net segmentation is shown in Figure 2.

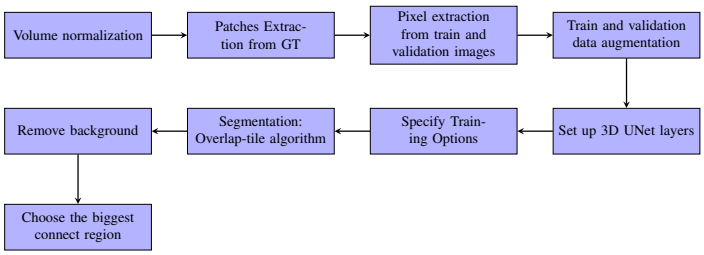


Figure 2: U-Net Process

2.3 Active Contours

Active contours algorithm, also known as snakes, consists of deforming the image domain and capture a desired feature through the constraint and image forces that pull it towards object contours and the internal forces resist deformation [3]. Two approaches to define the initial contour were used: the clustering results and the U-Net results. The optimal number of iteration was found by evaluating some cases with the metrics Dice, Jaccard and the BF (Boundary F1) contour matching score. Values between 150 and 325 with steps of 25 were evaluated. The segmentation results by defining the U-Net results as initial contours, went through the same morphological operations as the ones used to create the mask based on HU values. The results by defining the clustering results as initial contours did not, once there were no improvements. The full pipeline of the Active Contours segmentation is shown in Figure 3.

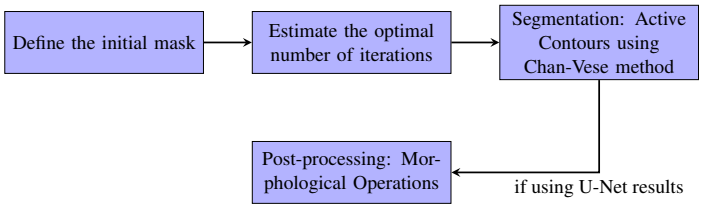


Figure 3: Active Contours Process

2.4 Graph Based

Graph-based image processing methods typically operate on pixel adjacency graphs. Adjacency graphs are graphs whose vertex set V is the set of image elements, and whose edge set E is given by an adjacency relation on the image elements. Initially, the optimal number of superpixels were estimated. For the HU mask, a label mask was computed by creating 400 superpixels on the image to segment and 350 for the anatomic mask. Then, a foreground mask and a background mask were created.These masks were created with three approaches: defining them manually by choosing ranges where the bladder is fully in (for the foreground mask) and where the bladder is fully out (for the background mask); using the clustering results (see Section 2.1) as the foreground mask and defining the background mask as a seed found based on ground truth coordinates; using the U-Net results (see Section 2.2) as the foreground mask and defining the background mask manually, once again. The full pipeline of the Graph Based segmentation is shown in Figure 4.

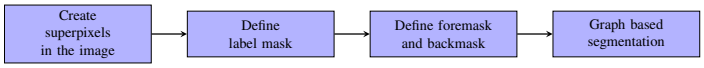


Figure 4: Graph Based Process

3 Dataset

The database was composed by 48 CT scans from 48 patients with prostate cancer, collected at the Institute of Oncology of Porto (IPO). Each CT has the structures manually segmented by specialists. Once most of CTs have more than one manual segmentation of the bladder, there are 71 bladders manually segmented. For some methods, it was needed to split the dataset into train, validation and test datasets. The train dataset consists of 19 patients, the validation has 4 and the remaining 24 went to test dataset. This leads to 37% of the structures to segment in the train dataset, 8% in validation and 55% in the test set.

4 Results

Segmentation illustrations are given in Figure 5, while the best results for each algorithm are given in Table 1.

Table 1: Algorithms comparison. "PP=MO" stands for morphological operations applied as pos processing

Mask	Algorithm	Dice	Jaccard	BF score	Precision	Recall
HU	Clustering (k=115 PP=MO)	0.2364 ± 0.1490	0.1421 ± 0.0972	0.8223 ± 0.1510	0.7269 ± 0.2146	0.9977 ± 0.2364
	U-Net	0.0617 ± 0.0318	0.0321 ± 0.0171	0.7030 ± 0.0580	0.5559 ± 0.0693	0.9648 ± 0.0617
	Active Contours (U-Net)	0.2603 ± 0.2280	0.1696 ± 0.1558	0.9686 ± 0.0099	0.9450 ± 0.0191	0.9935 ± 0.2603
	Graph Based (U-Net)	0.2903 ± 0.0975	0.1737 ± 0.0690	0.9675 ± 0.0084	0.9523 ± 0.0134	0.9832 ± 0.2903
Anatomic	Clustering (k=46 PP=MO)	0.2215 ± 0.1960	0.1381 ± 0.1248	0.6247 ± 0.4533	0.7024 ± 0.3209	0.6387 ± 0.2215
	U-Net	**0.8048 ± 0.2249**	0.7186 ± 0.2482	0.9981 ± 0.0036	0.9979 ± 0.0049	0.9984 ± 0.8048
	Active Contours (U-Net)	0.3069 ± 0.3112	0.2253 ± 0.2459	0.9654 ± 0.0081	0.9778 ± 0.0171	0.9933 ± 0.0059
	Graph Based (manual masks)	0.1990 ± 0.0863	0.1131 ± 0.0548	0.9640 ± 0.0112	0.9435 ± 0.0215	0.9856 ± 0.1990

5 Conclusion

Different approaches were developed with the aim of helping specialists to detect the bladder more effectively and quicker. Two pre procedures were implemented to detect the RoI. Using GT information, it was possible to create a mask smaller than the one based on the HU values of the bladder. Consequently, the results were better with this mask. Next, four segmentation methods were tested. The approach whose results stands out is using the U-Net algorithm to predict the label of each volume, followed by the application of morphological operations and the selection of the biggest connect region.

In the future, it is intended to use hyperparameter optimisation techniques to choose the optimal parameters in each algorithm, to obtain better results. Since the rectum is a region of interest in the context of prostate cancer, it is planned to implement the previous methods, with the aim of studying whether their performance is similar when segmenting the rectum.

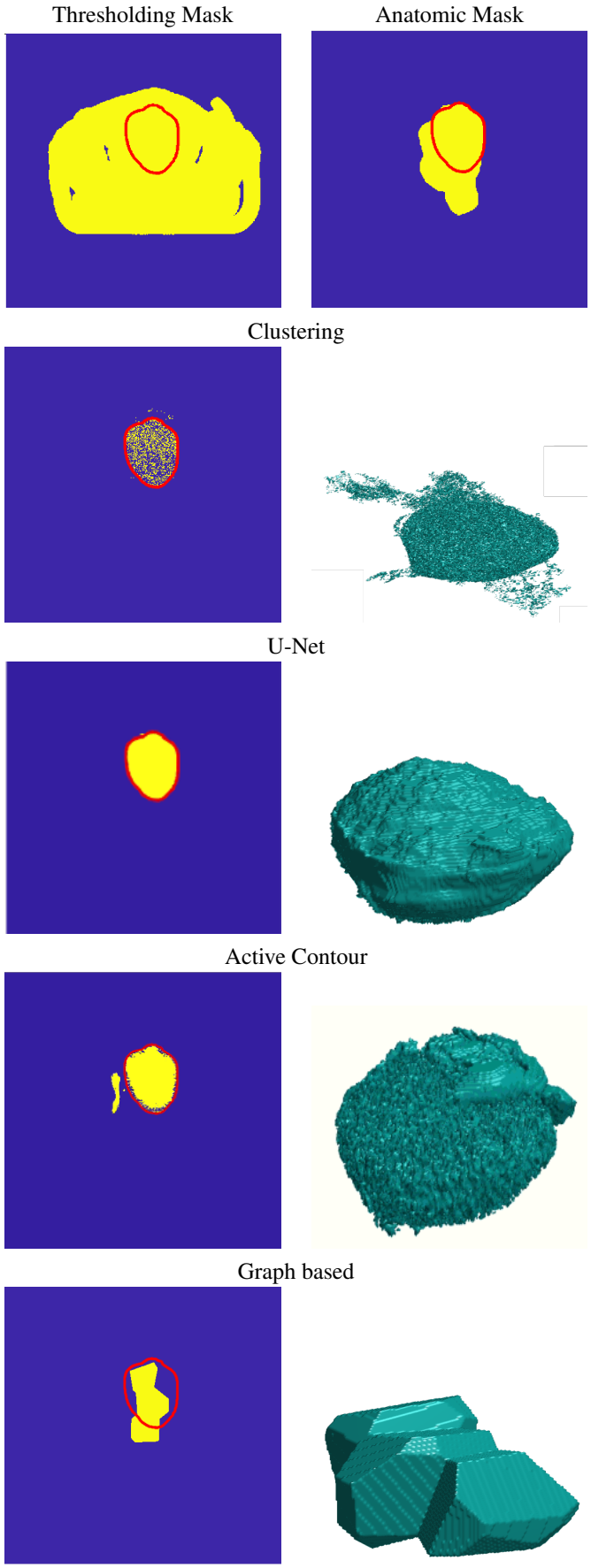


Figure 5: Selected segmentation results

References

[1] Neil G Burnet. Defining the tumour and target volumes for radiotherapy. *Cancer Imaging*, 4(2):153–161, 2004.

[2] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 424–432, 2016.

[3] Bing Li and Scott T Acton. Active Contour External Force Using Vector Field Convolution for Image Segmentation. *IEEE Transactions on Image Processing*, 16(8):2096–2106, 2007.

Segmentation of optic disc and cup for glaucoma analysis using cup-to-disc ratio

Alexandre Neto[1,2]
alexandre.hc.neto24@gmail.com
José Camera[2,3]
jcrcamara@hotmail.com
Sérgio Oliveira[1]
sergiooliveira4820@gmail.com
Ana Cláudia[1]
anaclaudia13ct@gmail.com
António Cunha[1,2]
acunha@utad.pt

[1]UTAD – Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal
[2]INESC TEC – Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência, Porto, Portugal
[3]UA – Universidade Aberta, Lisboa, Portugal

Abstract

Glaucoma is a silent disease that can lead to irreversible blindness. Early screening allows treating patients in time. For glaucoma screening, retinal images are very important since they enable the examination of initial glaucoma lesions, which typically begins with the cupping formation within the Optic Disc (OD). In clinical settings, practical indicators such as Cup-to-Disc Ratio (CDR) are frequently used to evaluate the presence and stage of glaucoma. Current Deep Learning (DL) methods can assist the glaucoma mass screening, lower the cost and allow it to be extended to larger populations. With DL methods in the OD and Optic Cup (OC) segmentation, it is possible to evaluate the presence of glaucoma in the patient more quickly based on cupping formation, using CDR. This work assesses the contribution of Multi-Class and Single-Class segmentation models with U-Net architecture to segment the OD and OC, and then evaluate glaucoma prediction based on different types of CDRs indicators. The segmentation of both OD and OC reach dice over 0.8 and IoU above 0.7. The CDRs were calculated for glaucoma assessment where was reached sensitivity above 0.8, specificity of 0.7, F1-Score around 0.7 and AUC above 0.85.

1 Introduction

The undetected glaucoma prevalence is as high as 90% in middle and low-income regions such as Asia and Africa, a consequence of the inadequate screening tools and strategies to detect these glaucomatous lesions [2]. Screening is a critical point, preventing this disease since glaucoma diagnosis at early stages gives the chance to stop its progression. Studies have revealed that morphological changes in the Optic Disc (OD) and in the Optic Cup (OC) indicate damage to the optic nerve, leading to glaucoma. The Cup-to-Disc Ratio (CDR) criterium, which is an indicator widely used by experts to detect glaucoma, measures the ratio between the cup and the disc [3], [4]. Computed-aided diagnosis solutions for screening glaucoma are in need in situations such as mass screening and medical care, even more in countries with a significant lack of qualified specialists [5]. Automatic segmentation approaches can be faster and more objective than humans [6]. However, most of the studies do not have a practical application of the segmentation results for glaucoma screening. The segmentation can help in mass screening by using the CDR indicator for glaucoma assessment.

1.1 Cup-to-Disc ratio

Glaucoma evolution can be assessed by measuring the ratio of the OD and OC. The CDR is a clinical method that compares the ratio of the cup to the disc. If the vertical CDR (VCDR, equation 1) or horizontal CDR (HCDR, equation 2) are more than 0.5, the eye can be considered a threat of abnormality. Otherwise, it is considered a normal eye [7]. Alternatively, considering the study of Andres Diaz Pinto [7], the assessment can be done through the area CDR (ACDR, equation 3) using the threshold of 0.3.

$$VCDR = \frac{V_{cup}}{V_{disc}} \quad (1) \qquad HCDR = \frac{H_{cup}}{H_{disc}} \quad (2) \qquad ACDR = \frac{A_{cup}}{A_{disc}} \quad (3)$$

This work will use segmentation models to compute the different types of CDRs for glaucoma assessment and evaluate the contributions for the practical application in glaucoma screening.

2 Methodology

In this work, glaucoma screening methods based on the segmentation approaches are evaluated. The pipeline used is described in Figure 1.

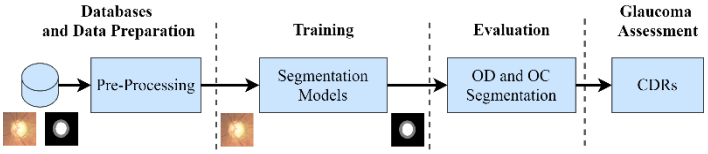


Figure 1: The pipeline of the work.

Retina and mask images from public databases are pre-processed and used to train Multi-Class (MC) and Single-Class (SC) segmentation models capable of segmenting the OD and OC in retinal images. The MC segmentation does the two tasks at once, while the SC has one model for OD segmentation and other for OC segmentation. The segmentations from the models are then evaluated and used to calculate the CDRs. In the end, the use of CDRs as indicators for glaucoma assessment is evaluated.

2.1 Databases and Data-Preparation

Three public datasets were used, namely the RIM-ONE r3 (85 healthy and 74 glaucoma images), DRISHTI-GS (31 healthy and 70 glaucoma images) and REFUGE database (360 healthy and 60 glaucoma images), with OD and OC masks and the respective labels. All three datasets were merged in a unique database (DB) where each image was centralised in the OD region and cropped into 512x512. Augmentations were applied (rotations, zooms and shifts variations) to avoid overfitting of the model. 70% of the images were used for training, 15% for validation and 15% for test.

2.2 Training

The pre-trained models (with ImageNet weights) selected were the Inception V3 and Inception ResNet V2 (for simplification, S1 and S2, respectively), based on related studies and high performance. First, the models were pre-trained for 20 epochs and after that fine-tuned for 100 epochs. The encoder weights are frozen for the pre-train and to fine-tune, the encoder layers are unfrozen to update all the weights. The learning rate started at 10^{-4}, using Adam optimiser and binary cross-entropy as the loss function.

2.3 Evaluation and Glaucoma Assessment

The metrics used for the evaluation of the segmentation model were the Intersection-Over-Union (IoU) and the dice coefficient. The glaucoma assessment using the different CDRs indicators is evaluated with sensitivity (Sen), specificity (Sep), F1-Score (F1) and the Area Under the ROC curve (AUC) by comparing to the true diagnosis.

	Masks GT			MC S1			MC S2			SC S1			SC S2		
	ACDR	HCDR	VCDR	ACDR	HCDR	VCDR	ACDR	HCDR	VCDR	ACDR	HCDR	VCDR	ACDR	HCDR	VCDR
Sen	0.82	0.86	0.86	0.82	0.82	0.89	0.86	0.93	0.93	0.69	0.73	0.85	0.85	0.85	**0.96**
Spe	0.81	0.64	0.75	**0.86**	0.64	0.79	0.83	0.61	0.74	**0.86**	0.64	0.74	0.84	0.56	0.70
F1	0.71	0.62	0.69	**0.75**	0.6	0.74	**0.75**	0.63	0.71	0.67	0.54	0.67	**0.75**	0.57	0.70
AUC	0.88	0.83	0.90	0.91	0.87	0.93	0.91	0.89	0.91	0.85	0.79	0.88	0.93	0.86	**0.96**

Table 2: Results of glaucoma classification with CDRs calculation for S1, S2 and Masks GT.

3 Results and Discussion

3.1 Segmentation results

Table 1 presents the results separately for the OD and OC segmentation using the MC and SC approach.

	OD		OC	
Method	**IoU**	**Dice**	**IoU**	**Dice**
MC S1	0.73 (±0.18)*	0.83 (±0.17)*	0.72 (±0.17)*	0.82 (±0.15)*
MC S2	0.72 (±0.18)*	0.82 (±0.17)*	0.71 (±0.18)*	0.82 (±0.16)*
SC S1	0.89 (±0.15)*	0.93 (±0.13)*	0.71 (±0.21)*	0.81 (±0.21)*
SC S2	**0.91 (±0.08)***	**0.95 (±0.05)***	**0.74 (±0.18)***	**0.83 (±0.16)***

* average (± standard deviation)

Table 1: Results for the OD and OC segmentation for our approach compared with the literature review results.

For the OD segmentation, the dice coefficient of the MC models is above 0.80 with 0.17 of standard deviation and for the SC models, the dice is over 0.9 with lower values of standard deviation. Since the SC models just focus on one element instead of the two, the results are better. The SC models show difficulties in the same samples that MC models had the worst results. The OC dice coefficient for both models in the MC approach is above 0.8 with 0.17 of standard deviation and for the SC models, the results are similar with dice above 0.8 as well. The calculation of CDRs is the core point since they can suggest the presence of glaucoma or not. The segmentation must be good enough so the use of CDR is reliable. Figure 2 illustrates some predictions for both models and the respective ground truth masks (Masks GT).

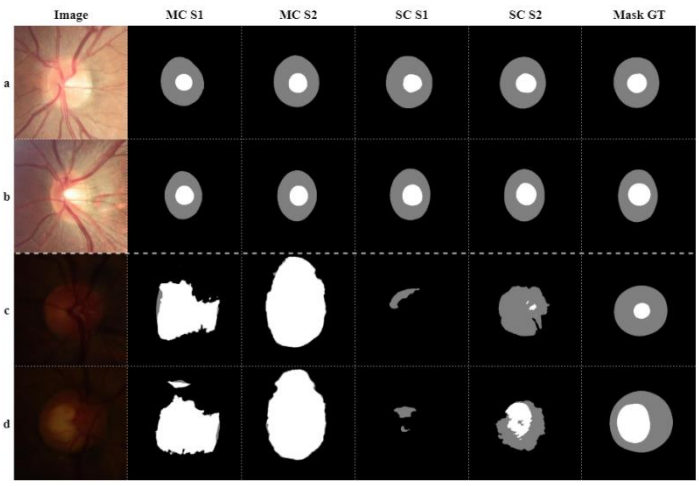


Figure 2: Examples of good results (a) and (b) and bad results (c) and (d), compared to the Mask GT of the respective images.

Figure 2 (a) and (b) represent good results that are close to the Mask GT. However, in (c) and (d), both models of the MC and SC approach reveal difficulties in correctly segmenting OD and OC, since the input images do not contain good enough quality/illumination to identify and segment the components. However, the bad results in all the approaches used were detected in low quality darker retinal images, which make the segmentation complicated even for a specialised professional. The same CDRs indicators are used to the Masks GT to have a direct comparison between the results from MC and SC approaches and the segmentation made by ophthalmologists, to see how reliable the segmentations made by S1 and S2 models are. Table 2 show the results of glaucoma assessment for both models in the MC and SC approaches that came close to, or even higher than the results using the Masks GT, which indicates that the segmentations are close to each other or at least provide close CDRs. Using the same CDR criterium for the segmentations of both models and the Masks GT shows similar values, supporting that the automatic segmentation can help to facilitate these tasks, providing similar results when evaluated using the CDRs indicators.

4 Conclusions

These models prove to be a good help on a subjective task that highly depends on the experience of the ophthalmologist and can contribute to expanding glaucoma screening to more people. The CDRs computed through the segmented masks were very close to the CDRs from the Masks GT. The model that reached better results overall for these tasks was Inception V3 as the backbone of the U-Net. The use of CDRs prove to be a good practical application of the segmentation since the results of glaucoma assessment using these indicators have high AUC results (lower AUC of 0.79 and higher AUC of 0.96).

Acknowledgments

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

References

[1] M. L. Claro, R. Veras, L. Santos, M. Frazão, A. Carvalho Filho, and D. Leite, “Métodos computacionais para segmentação do disco óptico em imagens de retina: uma revisão,” Rev. Bras. Comput. Apl., vol. 10, no. 2, pp. 29–43, 2018.

[2] S. MacIver, D. MacDonald, and C. L. Prokopich, “Screening , Diagnosis , and Management of Open Angle Glaucoma : An Evidence-Based Guideline for Canadian Optometrists,” Cjo, vol. 79, no. 1, pp. 1–72, 2017.

[3] J. Cheng et al., “Superpixel classification based optic cup segmentation,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8151 LNCS, no. PART 3, pp. 421–428, 2013.

[4] S. Sreng, N. Maneerat, K. Hamamoto, and K. Y. Win, “Deep learning for optic disc segmentation and glaucoma diagnosis on retinal images,” Appl. Sci., vol. 10, no. 14, 2020.

[5] A. Sevastopolsky, “Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network,” Pattern Recognit. Image Anal., vol. 27, no. 3, pp. 618–624, 2017.

[6] B. Al-Bander, B. M. Williams, W. Al-Nuaimy, M. A. Al-Taee, H. Pratt, and Y. Zheng, “Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis,” Symmetry (Basel)., vol. 10, no. 4, 2018.

[7] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, and A. Navea, “CNNs for automatic glaucoma assessment using fundus images: An extensive validation,” Biomed. Eng. Online, vol. 18, no. 1, pp. 1–19, 2019

Detection of polyps in colonoscopy images

Sara Nóbrega [1]
sarapiresnobrega@hotmail.com
José Ribeiro [1]
pedro.1101@outlook.pt
António Cunha [1,2]
acunha@utad.pt

[1] UTAD - University of Trás-os-Montes and Alto Douro, Vila Real, Portuga
[2] INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Portuga

Abstract

Polyps in the colon and rectum can be detected through colonoscopies. The manual detection and classification of polyps is a difficult process as well as tedious and prone to error. The performance of these processes is far from being perfect. Hence, this project aims to help physicians in polyp detection during colonoscopies using for that purpose Deep Learning methods. In this paper, multiple state-of-the-art Convolutional Neural Networks were tested as binary classifiers. All applied a transfer learning approach and achieved an average accuracy of 95,70% in the polyp detection task. To train, test, and evaluate the classifiers, multiple public datasets were used. The negative class was made of images that belonged to multiple classes (healthy tissue and pathologies), enabling the classifiers to distinguish polyps from other common findings in the lower gastrointestinal tract.

1 Introduction

Colorectal Cancer (CRC) is a major worldwide cause of mortality. The incidence of this cancer is superior to 9% and is the third most prevalent cancer. Additionally, it is the fourth most common cause of death, affecting both genders equally, and developed countries are the ones with the highest incidence rates [1].

Polyps (Figure 1) in the colon are initially benign but can become malignant. The remotion of these projections in the intestine lumen is the most effective form of treatment [2]. Colonoscopy is a screening method used to observe the lower part of the gastrointestinal tract. Using this technique, it is possible to detect polyps and remove them for analysis [3].

Figure 1: Example of polyp images collected during colonoscopy.

The survival rate is related to the stage of cancer at diagnosis. When detected at localised stages, the five-year survival rate of CRC is 90%, but for regional and distant metastatic cancer, the five-year survival rate is 70% and 10%, respectively [1].

Deep Learning (DL) includes methods such as Convolutional Neural Networks (CNNs) and has revolutionised conventional techniques since presents better results. CNNs are capable of extracting features automatically from the data and can be used to identify polyps in images [3].

In this work, different state-of-the-art models were trained to detect polyps in colonoscopy images. In the negative class were included multiple pathologies to increase the detection ability, helping them to distinguish polyps from other findings.

This paper is divided into four sections: Section 1 introduces the context and motivates the work; Section 2, the methodology is presented detailing data preparation and models training and evaluation; Section 3 presents and discusses the results achieved; Finally, Section 4 presents the conclusion and future works.

2 Methodology

Figure 2 shows the work pipeline.

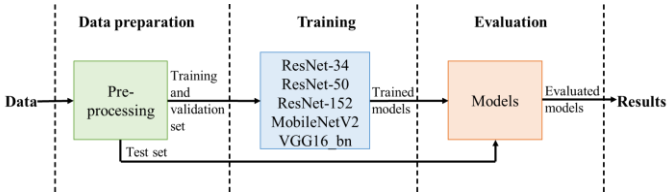


Figure 2: Work pipeline.

Three public datasets (Kvasir V2 [4], CVC-ClinicDB [5], and the ETIS-LaribPolypDB [6]) were merged, aiming to evaluate the different state-of-the-art architectures selected.

The totality of polyp images from all the datasets was used. Images from normal mucosa (normal pylorus, cecum, and z line), other pathologies (esophagitis and ulcerative colitis), and therapeutic intervention (dyed resection margins) were also selected from the Kvasir V2 dataset [4]. These images were included in the negative class. The main goal of inserting these images in the negative class is related to the discrimination ability of the classifiers: the models could distinguish polyps not only from normal mucosa but also from other malignancies.

In Table 1, it is possible to observe the train, validation, and test dataset splits used to develop and evaluate all the models.

Table 1. The number of polyp and non-polyp images used for the train, validation, and test sets.

Dataset	*Train*	*Validation*	*Test*	*Total*
Polyp	1.208	300	300	1.808
Non-polyp	1.600	700	700	3.000
Total	2.808	1.000	1.000	4.808

All the images were resized to 224x224 pixels. The alphanumeric characters and black margin present in some images were removed, by cropping them.

The architectures selected were: ResNet-34 [7], ResNet-50 [7], ResNet-152 [7], MobileNetV2 [8], and VGG16_bn [9]. The CNNs were trained for the binary classification, identifying polyps in colonoscopy images. A Transfer Learning (TL) approach was applied, and all the models were pre-trained on the ImageNet dataset [10], which were posteriorly fine-tuned, aiming to distinguish polyps from other possible findings in the colon.

The software libraries fast.ai and TensorFlow were used and all the experiments were implemented inside Google Colaboratory.

To evaluate the performance of the models, metrics based on the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) were selected. The selected metrics are the following:

$$\text{Accuracy} = (TP+TN) / (TP+FP+TN+FN) \quad (1)$$

$$\text{Recall} = TP / (TP+FN) \quad (2)$$

$$\text{Precision} = TP / (TP+FP) \quad (3)$$

$$\text{F1 Score} = (2*\text{Precision}*\text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

3 Results and Discussion

The Smith's One Cycle Policy [11] was used during training. This policy applies large, cyclical learning rates. This policy allows to train quicker the models, reduces overfitting, and makes it possible to reach higher accuracy.

The performance of the models in the test set can be consulted in Table 2.

Table 2. Results obtained by the state-of-the-art models in the test set.

Model	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1 Score (%)*
ResNet-34 [7]	94,90	92,80	90,00	91,38
ResNet-50 [7]	96,30	94,60	93,00	93,79
ResNet-152 [7]	**96,60**	94,30	**94,30**	**94,30**
MobileNetV2 [8]	95,80	93,30	92,70	93,00
VGG16_bn [9]	94,80	**96,70**	85,70	90,87

Analysing the results present in Table 2, it is possible to affirm that all the models had a similar performance. Additionally, observing the same table it is possible to conclude that no matter the depth of the ResNet architecture analysed, the results are similar but with slightly better performance in the ones with more depth (ResNet-50 [7] and ResNet-152 [7]). Although, VGG16_bn [9] which is the model with the smallest depth investigated, it achieved similar results when compared to ResNet-152 that was the model with the highest depth.

Since deeper architectures only produce a minimal increase in performance, their use is not justifiable considering the computational cost required.

The review article [3] reviewed multiple polyp detection studies. Notice that the results achieved by all the models in this work are like the ones produced by the state-of-the-art models considered.

Figure 3 represents a TP, TN, FP, and FN produced by the ResNet-50 [7] model. Observing predictions in the test set made by the ResNet-50 [7] model, it was possible to observe in some of the FPs, normal colonic mucosa textures similar to polyps. The FP present in Figure 3 is representative of this situation.

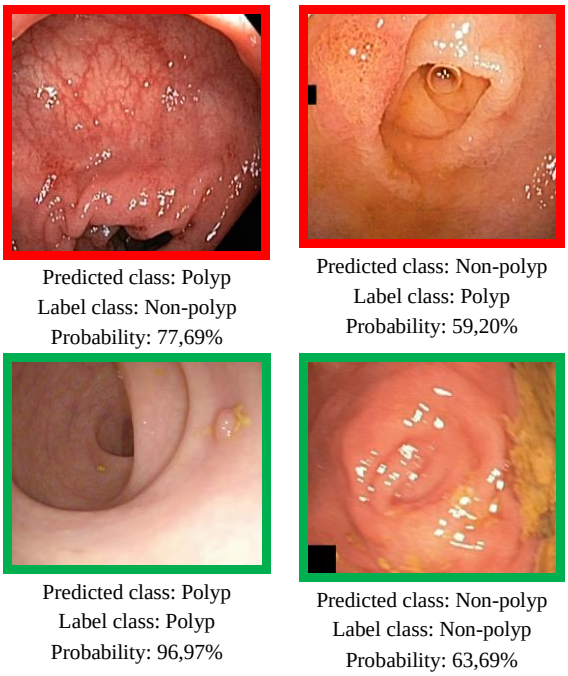


Figure 3: Images and respective predictions made by the ResNet-50 [7] model. The "Probability" is the confidence of the model in the prediction.

Observing Figure 3, despite the polyp present in the TP has a small size, the model was capable of detecting it. As for the TN, it is possible to observe some motion blur associated with the endoscope movement, but the model could recognise that there was no polyp in the image.

4 Conclusion

This study evaluated different state-of-the-art models to detect polyps in colonoscopy images. On average, the models reached an accuracy of 95,70%, which indicates that all the models were capable of correctly classify a high percentage of images.

Since multiple architectures with different depths were selected, it was possible to observe the effect of the number of layers produced in the model's performance. The results demonstrate that the different state-of-the-art architectures reached similar performances. Hence, architectures with a smaller depth are interesting approaches since they need smaller computational resources.

Concerning ResNet-50 [7], after carefully examining the results in the test images, it was possible to observe that the model mistook some normal intestinal structures, such as folds, as polyps. These structures also have a volume that stands out in of remaining parts of the image. Therefore, to reduce the FPs, it could be useful to include in the dataset more images representative of these normal intestinal structures, enabling the model to learn how to distinguish polyps from folds. Besides, future work could include the use of post-hoc explainable artificial intelligence techniques that can be used to observe which image information is being considered to make predictions.

DL techniques, such as the ones explored in this work, could improve the detection of polyps rate, selecting images with this pathology to be analysed by specialists in the field.

Acknowledgments

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

References

[1] Haggar, F. A., & Boushey, R. P. (2009). Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. Clinics in colon and rectal surgery, 22(4), 191.

[2] Marks, J. W. (2019, December 17). 8 Colon Polyps Symptoms, Pictures, Types, Causes, Treatment. MedicineNet. https://www.medicinenet.com/colon_polyps/article.htm

[3] Sánchez-Peralta, L. F., Bote-Curiel, L., Picón, A., Sánchez-Margallo, F. M., & Pagador, J. B. (2020). Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. Artificial intelligence in medicine, 101923.

[4] Pogorelov, K., Randel, K. R., Griwodz, C., Eskeland, S. L., de Lange, T., Johansen, D., ... & Halvorsen, P. (2017, June). Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In Proceedings of the 8th ACM on Multimedia Systems Conference (pp. 164-169).

[5] Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., & Vilariño, F. (2015). WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics, 43, 99-111.

[6] Angermann, Q., Histace, A., Romain, O., Dray, X., Pinna, A., & Granado, B. (2015). Smart videocapsule for early diagnosis of colorectal cancer: toward embedded image analysis. In Computational Intelligence in Digital and Network Designs and Applications (pp. 325-350). Springer, Cham.

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[8] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[9] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[10] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[11] Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820.

Preliminary Study on the Impact of Attention Mechanisms for Medical Image Classification

Tiago Gonçalves[1,2]
tiago.f.goncalves@inesctec.pt

Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt

[1]Faculdade de Engenharia
Universidade do Porto
Porto, Portugal

[2]INESC TEC
Porto, Portugal

Abstract

Despite their high performance, deep learning algorithms still work as black boxes and are not capable of explaining their predictions in a human-understandable manner, thus leading to a lack of transparency which may jeopardise the acceptance of these technologies by the healthcare community. Therefore, the topic of explainable artificial intelligence (xAI) appeared to address this issue. There are three main approaches to xAI: pre-, in- and post-model methods. In medical images, important information is generally spatially constricted. Hence, to ensure that models focus on the important parts of the images and learn relevant features, several attention mechanisms have been proposed and demonstrated increased performances. This work proposes a comparative study of the application of different attention mechanisms in deep neural networks and the evaluation of their impact on the performance of the models and the quality of the learned features.

1 Introduction

The democratised access to data and the increase of the availability of computational power allowed deep learning (DL) methodologies to achieve nearly-human performances in several areas of science, business and government. The popularity and success of DL in computer vision is mainly due to the introduction of convolutional neural networks (CNNs), which are designed to process unstructured data (*e.g.*, images) [6]. In medical image classification, the main task is to output a diagnosis (*e.g.*, presence or absence of a disease) based on one or more input images. Given the high predictive performance rates of CNNs in other computer vision tasks (*e.g.*, natural image recognition), the application of DL algorithms in medical image classification occurred almost naturally.

2 Explainable Artificial Intelligence

Despite the high performances achieved by DL-based algorithms, their transition into real-world applications is not trivial, due to their complexity (*i.e.*, high-number of parameters) and their black-box behaviour, which may jeopardise their acceptance by the clinical community. Therefore, the topic of explainable artificial intelligence (xAI) appeared intending to contribute to a more transparent AI. Although there is no clear distinction between explainability and interpretability, one may think of these as a three-stage process [3]: pre-model methods focus on understanding the data distribution before building the model, through exploratory data analysis; in-model methods seek to integrate interpretability inside the model (*e.g.*, models based in rules, models based in cases, the use of regularisation techniques during training to obtain sparser or monotonic models); post-model methods are related to a posterior analysis of the model predictions (*e.g.*, using the gradient information to identify the areas of the image that mostly contribute to the final decision, inserting a perturbation and observing the prediction, inverting the representations back to the input pixel space or connecting the representations to semantic concepts). In healthcare applications, it is fundamental to assess the quality of these explanations, for the sake of transparency, ethics and fairness [8].

3 Attention Mechanisms

The intuition behind the application of attention mechanisms in DL algorithms is inspired by the field of psychology, according to which humans tend to selectively concentrate on a part of the information. For instance, the human visual system tends to selectively focus on specific parts of an image while ignoring others. Following this rationale, it is recognised that in AI systems, some parts of the inputs may be more relevant than others (*e.g.*, in automatic translation systems, only a subset of words is relevant). The use of attention was initially proposed in [1] for the task of neural machine translation. Recently, a CNN with a multi-level dual-attention mechanism (MLDAM) has been proposed for macular optical coherence tomography classification [7]. The main novelty of this work in the context of medical image classification is the joint application of a *self-attention* and a *multi-level attention* mechanisms that allow the network to learn relevant features in coarser as well as finer sub-spaces. Regarding the impact of the application of attention mechanisms in the interpretability of the DL algorithms, we point to the work proposed by [2], which approaches the field of interpretability through an analysis of the saliency maps produced by the gradient-weighted class activation mapping (Grad-CAM) [9].

4 Data

We decided to perform experiments on two different use-cases using medical images: breast cancer detection in mammography (CBIS-DDSM data set) and pathology detection in chest X-ray (MIMIC-CXR data set). Each data set contains images of two different classes (binary classification): normal (*i.e.*, without lesion or pathology) and abnormal (*i.e.*, with lesion or pathology).

5 Implementation

We performed a comparative study using three state-of-the-art pre-trained deep learning models as backbones: VGG-16 [10], ResNet-50 [4] and DenseNet-121 [5]. To assess the influence of the use of attention mechanisms, we adapted the MLDAM architecture described in [7] for each of the backbones. We performed experiments with four use-cases: **baseline** (*i.e.*, only the backbone is trained), **baseline with data augmentation** (*i.e.*, the backbone is trained with data augmentation strategies), **baseline and MLDAM** (*i.e.*, the backbone with MLDAM is trained), **baseline and MLDAM with data augmentation** (*i.e.*, the backbone with MLDAM is trained with data augmentation strategies). All the images are resized to the final size of 224×224 and a z-normalisation is applied to each RGB channel. The data augmentation strategy employed in this work is composed of several random rotations, random translations, random scaling, and random horizontal flips. Each model is trained for a maximum of 300 epochs, with binary cross-entropy as the loss function and Adaptive Moment Estimation (Adam) with learning rate 1×10^{-4} as the optimisation algorithm. The batch size varied from 1 to 4, depending on the available GPU memory. We save the best model's parameters in both training and validation sets according to the value of the loss. We tested all the trained models (using the best weights in both training and validation sets) in the test set of each database and computed the accuracy, precision, recall and F1-score. We generated saliency maps [11] for the positive and negative samples of the test set that were correctly predicted by all the use-cases related to all backbones, to assure a fair intra- and inter-comparison. It is important to note that these saliency maps were generated using the models loaded with the best weights in the validation set of each database.

6 Results and Discussion

An extended version of these results is publicly available in a GitHub repository[1]. Table 1 and Table 2 present the accuracy results obtained for the test set of the CBIS-DDSM and MIMIC-CXR, respectively. In both cases, we can observe that the models' predictive performance does not suffer abrupt changes. Figure 1 and Figure 2 present examples of saliency maps obtained for images with label "0" of the CBIS-DDSM

[1] https://github.com/TiagoFilipeSousaGoncalves/attention-mechanisms-healthcare/blob/main/reports/Report.pdf

Table 1: Accuracy results obtained for the test set of the CBIS-DDSM data set: (a) - Baseline, (b) - Baseline with Data Augmentation, (c) - Baseline and MLDAM, (d) - Baseline and MLDAM with Data Augmentation.

Model	Weights	(a)	(b)	(c)	(d)
DenseNet-121	Training	0.6650	0.6142	0.6210	0.5871
	Validation	0.6108	0.5584	0.6396	0.6125
ResNet-50	Training	0.6514	0.6396	-	0.5973
	Validation	0.5956	0.6176	0.5939	0.5854
VGG-16	Training	0.6650	0.6074	0.5939	0.5854
	Validation	0.6244	0.6514	0.6210	0.6041

Table 2: Accuracy results obtained for the test set of the MIMIC-CXR data set: (a) - Baseline, (b) - Baseline with Data Augmentation, (c) - Baseline and MLDAM, (d) - Baseline and MLDAM with Data Augmentation.

Model	Weights	(a)	(b)	(c)	(d)
DenseNet-121	Training	0.8451	0.8312	0.8349	0.8386
	Validation	0.8629	0.8498	0.8535	0.8666
ResNet-50	Training	0.8340	0.8424	0.8470	0.8386
	Validation	0.8535	0.8694	0.8563	0.8414
VGG-16	Training	0.8507	0.8330	0.8293	0.8461
	Validation	0.8629	0.8731	0.8535	0.8647

and MIMIC-CXR, respectively. Taking into account the reported effects of the use of attention mechanisms in the quality of the features learned during training, we were expecting the saliency maps to highlight clear differences between the baseline (with or without data augmentation) and the baseline with an attention mechanism (with or without data augmentation). However, the saliency maps obtained suggest that the behaviour of the models is either similar or completely disparate, without apparent meaning. This lack of consistency makes it difficult to relate direct benefits to the use of attention mechanisms on the properties of post-model explanation methods.

7 Conclusions and Future Work

Our experiments did not present conclusive results on the impact of attention mechanisms in two healthcare use-cases, using three different state-of-the-art backbones. Hence, further work should be devoted to: 1) the development of new experiences with different data processing and augmentation strategies, since it is not clear if these steps are harming the performance of the models; 2) the design of different attention mechanisms that capture features from different scales or levels, since, to the authors' knowledge, there is not a clear pipeline on which are the best scales or levels that should be incorporated in an MLDAM module; 3) generate saliency maps with other methods to see if the results that we obtained are dependent of the post-model interpretability method or not; 4) experiment different state-of-the-art backbones to see if their behaviour differs from the ones we used in this work; 5) try different data sets and different tasks to assess if results are data or task-dependent.

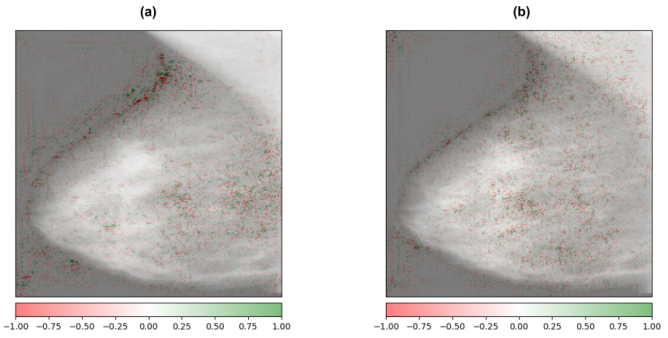


Figure 1: Examples of saliency maps obtained for an image with label "0" of the CBIS-DDSM data set, using the DenseNet-121 backbone model: (a) - Baseline, (b) - Baseline and MLDAM.

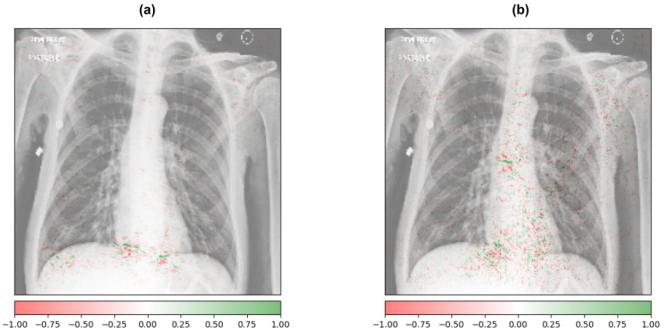


Figure 2: Examples of saliency maps obtained for an image with label "0" of the MIMIC-CXR data set, using the DenseNet-121 backbone model: (a) - Baseline, (b) - Baseline and MLDAM.

Acknowledgements

This work was partially funded by the Project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership and the PhD grant "2020.06434.BD".

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Cunjian Chen and Arun Ross. An explainable attention-guided iris presentation attack detector. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 97–106.

[3] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*, March 2017. arXiv: 1702.08608.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[6] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature14539.

[7] Sapna S Mishra, Bappaditya Mandal, and Niladri B Puhan. Multi-level dual-attention based cnn for macular optical coherence tomography classification. *IEEE Signal Processing Letters*, 26(12):1793–1797, 2019.

[8] Cynthia Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *arXiv:1811.10154 [cs, stat]*, September 2019. arXiv: 1811.10154.

[9] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[11] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

False-positives attenuation of automatically detected hotspots on bone scintigraphy images using image analysis techniques

Laura Providência
lauraprovid@hotmail.com
Inês Domingues
inesdomingues@gmail.com
João Santos
joao.santos@ipoporto.min-saude.pt

Faculdade de Ciências da Universidade do Porto
Portugal
Medical Physics, Radiobiology and Radiation Protection Group, IPO Porto Research Centre (CI-IPOP), Portugal
Instituto de Ciência Biomédicas Abel Salazar
Porto, Portugal

Abstract

Prostate cancer (PCa) is the second most diagnosed cancer in men. Patients with PCa often develop metastases, with more than 80% of this metastases occurring in bone. The most common imaging technique used for screening, diagnosis and follow-up of disease evolution is bone scintigraphy, due to its high sensitivity and widespread availability at nuclear medicine facilities. To date, the assessment of bone scans relies solely on the interpretation of an expert physician who visually assesses the scan. Besides this being a time consuming task, it is also subjective, as there is no absolute criteria neither to identify bone metastases neither to quantify them by a straightforward and universally accepted procedure. In this paper, an algorithm which uses image analysis techniques for the false positives reduction of automatically detected hotspots in bone scintigraphy images is proposed. The final algorithm correctly identified 30% of the non-malignant hotspots from the data set as false-positives, and decreased the number of false positive images per image in 31%.

1 Introduction

According to the World Health Organization, prostate cancer (PCa) is the second most commonly diagnosed cancer in men, accounting for more than 1.4 million new cases and more than 375 000 deaths worldwide in 2020. Patients with advanced prostate cancer often develop metastases, with the bone being the most frequent site for metastatic growth. Currently there is no cure for metastatic prostate cancer, but it can often still be treated to slow down its growth. A precise detection and up-take quantification of bone metastases is essential to provide the physicians the accurate staging they require to choose the appropriate treatment for an individual patient, to monitor the evolution of the disease, and to evaluate the treatment efficiency.

The most common diagnostic procedure used for screening, assessment of treatment and follow-up of patients with bone metastases is whole-body bone scintigraphy (BS) [1]. In a bone scintigraphy, also known as bone scan, a simultaneous image of the anterior (AP) and posterior (PA) views of the patient is obtained. The scans will reveal brighter areas, which indicate an increased rate of bone metabolic activity, such as abnormal growth caused by metastases. These areas are referred to as hotspots, and may indicate not only the presence of bone metastases, but also other conditions such as trauma, micro- arthritis, benign degeneration, or bone infections. The biggest disadvantage in the use of bone scintigraphy to detect bone metastases is, therefore, its low specificity. Because it evaluates the distribution of active bone formation in the skeleton and identifies the sites where metabolic reactions are occurring, it detects several suspicious uptakes of non-metastatic origin, which lead to high a false-positive (FP) rate of BS to detect bone metastases.

This work aims to create an algorithm capable of identifying hotspots with no metastatic origin (FPs) in bone scintigraphy images. Such an algorithm could be used on its own or as a complement to a hotspots classification algorithm (such as the one proposed in [5]) to build a software capable of identifying bone metastases, providing the physician with a fast, precise and reliable tool to quantify bone scans and evaluate disease progression and response to treatment.

2 Materials and Methods

2.1 Database

The database consists of 30 bone scintigraphy images from 22 patients with prostate cancer with suspected bone metastatic disease. The equipment used for scanning patients was either a *Millennium MG* (GE Medical Systems), which digitally record anterior and posterior scans with a resolution of 1024×256 pixels, or a *BrightView* (Philips Healthcare), which digitally records anterior and posterior scans with a resolution of 1024×512 pixels. The pixel depth (maximum number of counts which could be stored in a pixel) is 16-bits for every image. For each bone scan, a medical report describing the condition of the patient in question written by a nuclear medicine physician is available. All data was provided by Instituto Português de Oncologia do Porto Francisco Gentil (IPO Porto). The data was collected and held anonymously and the developed algorithms did not contain information concerning the patients, but rather information extracted from the data during the algorithm development. This project was authorized by IPO-Porto Healthcare Ethics Committee.

The scans were equally divided into one of three possible categories: (i) *healthy*, if no suspicious bone uptake was detected, (ii) *benign*, if bone hotspots with no metastatic origin were present or (iii) *malignant*, if bone metastases existed in the scan. Hotspots were extracted from each bone scan using a technique based on the approach proposed in [3] (a thoroughly explanation of this algorithm can be found in [5]). Table 1 summarizes the available database, including the number of bone scans and hotspots extracted per category. It is important to point out that images from the malignant category can also present benign and healthy hotspots, just like images from the benign category can also present healthy hotspots.

Table 1: Database summary. The database consists of a total of 1311 hotspots extracted from 30 bone scans.

Bone scan type	No of bone scans	No of hotspots
Healthy	10	138
Benign	10	255
Malignant	10	918
Total	**30**	**1311**

2.2 Methodology

The algorithm used to extracted the hotspots was programmed to find brighter regions in the scans, which can represent anything from healthy physiological processes, benign lesions to metastases. This meant that, even though the algorithm successfully detected the metastases, it would also detect a considerable amount of hotspots not related to bone metastases (average of 32 false positive detections per image). Since the patient condition is determined through the assessment of the malignant bone lesions, the number of FP detections should be reduced. This was achieved through the development of an algorithm that uses image analysis techniques to identify FP detections.

2.3 Attenuation of false-positives

There are hotspots that, due to some specific characteristic that they present, can be easily identified as FPs. Using solely image analysis techniques, they can be detected and removed. These include:

- ***Hotspots found in certain anatomical regions***. There are certain anatomical regions where FP hotspots are commonly detected. For example, increased radiotracer uptake is common in urine, and therefore a noticeable hotspot in the bladder is almost always seen. Another common place for a hotspot to appear is in the hand, as this is usually the place trough which the radioisotope is injected.

To remove these hotspots, an atlas-based method for the anatomical segmentation of the bone scans was developed, which allowed the automatic localisation of the hotspots into one of eighteen different body regions. Hotspots belonging to the bladder, hands and feet were removed, as well as hotspots located outside of the body, as they corresponded to urine-collection bags.

- ***Symmetrical hotspots***: The appearance of symmetrical hotspots in bone scans is usually related to normal physiological processes. To find them, an algorithm to detect the symmetry axis of a patient in a bone scintigraphy image was used. The code used for the identification of the symmetry axis was developed by [2] and is available at [4]. For being considered symmetric, two hotspots had to verify the following conditions:
  - The absolute difference between the perpendicular distance from the hotspot centroids to the symmetry axis could not exceed a certain threshold, T_{dist} (set to 7.5 pixels);
  - The hotspots must lay on opposites side of the axis;
  - The absolute difference between the y-coordinates of the hotspot's centroids could not exceed a certain threshold, T_y (set to 5 pixels);
  - The absolute difference between the areas of the hotspots could not exceed a certain threshold, T_{area} (set to 30% of the area of one of the hotspots).

3 Results

In this section, the results are reported. The ground truth was obtained by manually labelling the hotspots detected in the bone scans from the data set as "0" (non-malignant) if they were considered to be FPs and "1" (malignant) if they were metastases. These labels were obtained manually according to the respective medical reports. The here proposed algorithm for FP attenuation was applied to each image from the data set, and the hotspots identified as FP detections were labelled as non-malignant, while the remaining hotspots were considered to be malignant. The predicted labels were compared with the true labels and the sensitivity, specificity, false negative rate (FNR), and false positive detections per image (FPPI) were calculated (Table 2). Figures 1 and 2 and illustrate the algorithm applied to two bone scintigraphy images from the data set and the confusion matrix of the final algorithm, respectively.

3.1 Discussion

With this algorithm, a high sensitivity was achieved (89%). This was expected, as the goal was to remove FP detections without losing any of the true-positive ones (metastases). Even so, the fact that this value was not 100% shows that a few metastases were lost, meaning that some malignant hotspot fell under the symmetry/region conditions. In the medical context, having a sensitivity inferior to 100% means that some metastases are falsely being classified as healthy hotspot, which would have significant impact on the patient's health. The algorithm sensitivity should therefore be improved before being used in the clinical practice. The specificity score shows that with this algorithm 30% of the non-malignant hotspots were correctly identified as FPs. Comparing with the score from the detection algorithm, the FPPI decreased by 30%. This value could be increased if a broader range of values for the symmetry conditions were considered; the reason why did was not done is because it would come at the cost of more malignant hotspots being wrongly considered FPs, specially in patients with high density of metastases.

4 Conclusions

An algorithm for the attenuation of FP detections in bone scintigraphy images with PCa was proposed. It can be used in combination with computer-assisted detection approaches to develop an automatic algorithm capable of quantifying whole-body bone scans, which would be extremely useful in the medical community. Improvements on the algorithm include finding symmetry/region conditions that keep the specificity at a maximum value (more ability to identify FPs), under the condition that no metastases are being lost (sensitivity = 100%). This could be achieved, for example, by building an optimisation algorithm. The algorithm also still leaves many FPs undetected. In the future, the algorithm proposed in [5] will also be applied to the images so that this number can be even more reduced, and a more accurate detection of the metastases can be done.

Table 2: Results of the algorithm when removing (i) only symmetrical hotspots, (ii) only hotspots found in certain anatomical regions and (iii) symmetrical hotspots and hotspots found in certain anatomical regions.

	(i) Symmetry	(ii) Region	(iii) Total
Sensitivity	0.93	0.95	0.89
Specificity	0.10	0.24	0.31
FNR	0.07	0.04	0.11
TN	87	229	296
FFPI	**29**	**24**	**22**

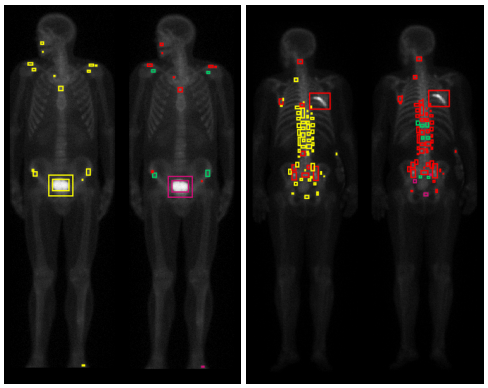


Figure 1: FP attenuation algorithm applied to two bone scintigraphy images from the data set, one from the healthy category (left) and other from the malignant category (right). For each scintigraphy, the image on the left represents the ground truth (GT), and the image on the right the classification according to the proposed algorithm. In the GT images, yellow and red bounding boxes represent non-malignant hotspots and metastases, respectively. In the images classified with the attenuation algorithm, pink and green bounding boxes represent detections considered to be FPs due to their anatomical position and condition of symmetry, respectively; red bounding boxes represent detections that were not considered to be FP.

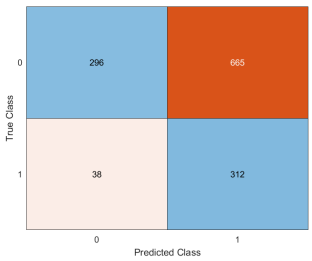


Figure 2: Confusion matrix

References

[1] I. Brenner A, J. Koshy, J. Morey, C. Lin, and J. DiPoce. The bone scan. *Seminars in Nuclear Medicine*, 42, 2012. Planar Imaging in the Age of SPECT.

[2] M. Cicconet, D. G. C. Hildebrand, and H. Elliott. Finding mirror symmetry via registration and optimal symmetric pairwise assignment of curves. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[3] I. Domingues and J. S. Cardoso. Using Bayesian surprise to detect calcifications in mammogram images. In *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014.

[4] GitHub. Symmetryviaregistration. URL `https://github.com/cicconet/SymmetryViaRegistration`. Accessed on 16.07.2021.

[5] L. Providência, I. Domingues, and J. Santos. An iterative algorithm for semisupervised classification of hotspots on bone scintigraphies of patients with prostate cancer. *Journal of Imaging*, 7(8), 2021.

Analysis of classification tradeoff in deep learning for gastric cancer detection

Gabriel Lima
FCUP,
INESC TEC,
Porto, PT

Miguel Coimbra
FCUP,
INESC TEC,
Porto, PT

Francesco Renna
Instituto de Telecomunicações,
INESC TEC,
Porto, PT

Abstract

This study aimed to build CNN models capable of classifying upper endoscopy images, to determine the stage of infection in the development of a gastric cancer. Two different problems were covered.

A first one with a smaller number of categorical classes and a lower degree of detail to perform a Macro classification. A second one had a Micro approach consisting of a larger number of classes, corresponding to each stage of infection of a gastric cancer in the Correa cascade. Three public datasets were used to built the dataset that served as input for the classification tasks. From the different CNN models, DenseNet169 achieved 0.72 for accuracy performance metric in the Micro approach classification task. The CNN models built for this study are capable of identifying the stage of a gastric lesion in the moment of an upper endoscopy. However, there is a trade-off between the specificity of the classification and the performance of the models. The tradeoff between detail in the definition of lesion classes and classification performance has been explored. Results from the application of Grad CAMs to the trained models shown that the proposed CNN architectures base their classification output on the extraction of physiologically relevant image features.

1 Introduction

Gastric cancer represents 7% of the world's cancers and 9% of the worldwide cancer related deaths. It is the fifth most frequent cancer type and the third leading cause of death from cancer, with approximately 950000 new cases and 783000 deaths in 2018 [1].

Performing diagnosis over lesions seen in the gastrointestinal tract during an upper endoscopy can be a difficult task, due to the wide variety of lesions that can happen in all the structures that are observable during the exam. A computer-aided diagnosis system helps medical personal in avoiding misdiagnosis with their predictions. The scope of this study was creating a deep learning algorithm for lesion classification on the subject of gastric and esophageal cancer based on images retrieved from high endoscopy exams. Two different multi class classification tasks will be explored. One will include more broad concepts, leading to more comprehensive class designations (healthy tissue, pre-cancerous lesions, cancerous lesions). The other will be more in-depth, with a separation of the data in a larger number of classes, where the class names are more specific lesions coming from the Correa's cascade for the development of gastric cancer (and not just stages).

In contrast with related work previously presented, in this paper we consider: 1) A classification problem definition for different cancerous and precancerous lesions, developed according to a clinically accepted criterion for their characterisation; 2) The impact of a different level of detail in the lesion characterisation versus classification performance. The algorithm of choice will be on the Artificial Neural Networks topic, more specifically on Convolutional Neural Networks (CNN). From the existent CNN architectures, several different frameworks were chosen. The main strategy that will be used is a transfer learning strategy, which will allow for the models to have already learned at the time of the training stage, by previously performing other classification tasks. Furthermore, one other tool considered to analyse the functioning and performance of the models will be the Grad CAMs. With this technique, it will be possible to understand if the models are performing the classifications based on a region of interest in the image or if their prediction was done using features that do not represent the focus of the lesions.

2 Related work

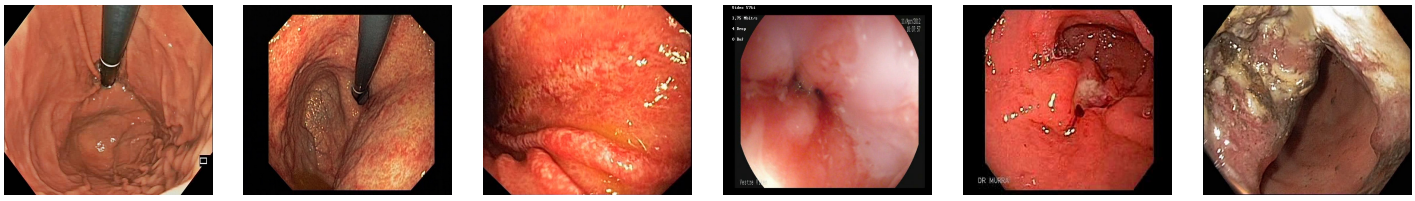
Currently, researchers have already performed academic studies whose main target was to use artificial intelligence and deep learning techniques to detect the appearance of stomach cancers. Cho et al. [2] explored the classification of stomach neoplasms through endoscopy images using different convolutional neural networks architectures. The data used in this project were endoscopic white-light images of pathologically confirmed gastric lesions. The dataset had a total of 5017 images from 1269 patients. The dataset in this paper was divided in 5 different classes. These 5 classes were advanced gastric cancer, early gastric cancer, high grade dysplasia, low grade dysplasia, and nonneoplasm. Regarding the CNN models, three models were built: Inception-v4 (IV4), Resnet-152 (RN152) and Inception-Resnet-v2 (IRV2), pretrained models in the ImageNet Dataset using transfer learning were adopted. The IRV2 had the best performance with 0.85 of accuracy for the 5-class classification problem.

Itoh et al. [4] follows a different approach studying the presence of Helicobacter-Pylori (HP) related infections. The purpose of the article was to create a CNN, capable of detecting and diagnosing an early infection caused by the presence of HP. The input also came from upper endoscopy images of patients. The training set images were obtained from the lesser curvature of the stomach in a total of 596 images. The authors also selected images of the selected curvature of the stomach which meant a higher sensitivity characteristic of this area and the diagnosis of HP infection was simplified. The model built was a CNN with resource to the GoogLeNet Deep CNN for standard object recognition. Transfer learning was once again used for the learning of the training data. This paper's task consisted of a binary classification task whose main goal was to detect the presence of HP-related lesions. Considering the intervals of classification for positive and negative, the sensitivity was 0.87, the specificity 0.87 and the AUC (Area Under Curve - performance metric for the classification problems at various threshold settings) was 0.96.

One last study concerned a classification problem related to anatomical landmarks detection, also using images from upper endoscopy exams. He et al. [3] performed a 12 class classification task, using upper endoscopy images. This study intended to build several models capable of classifying images from a dataset with 3704 upper endoscopy exams. This data was further divided by the authors in 4 different datasets. There were several CNN architectures chosen by the authors to perform the classification task, with the authors having performed fine-tunning of the models and also performed transfer learning to build them. One of the CNN models built was the Densenet-121. This was the best model in two of the datasets, with an accuracy score of 0.91 and 0.88, making it the best model. Finally, this study aimed to implement an anatomical site classification method in upper endoscopy images. The results showed it was possible and effective to carry out such study in a small number of images from the gastrointestinal tract.

3 Materials and dataset construction

According to what was seen regarding the stages of the development of a gastric cancer and the Correa's cascade, the classes chosen for the Macro class problem were: Healthy (HE), Precancerous (PRC) and Cancerous (CAN). Afterwards, these classes were then separated (apart from HE) in more extensive sets of classes that would go into further detail clinically speaking. Regarding the Micro Class problem the following classes were created: Healthy (HE), Atrophic Gastritis (AG), Intestinal Metaplasia (IM), Barrett's Esophagus (BE), Early Gastric Cancer (EGC) and



(HE) (AG) (IM) (BE) (EGC) (AGC)

Figure 1: Example of each class in the dataset of the Micro class classification problem.

Advanced Gastric Cancer (AGC). An example of each class can be seen in Figure 1. The datasets from which the images were collected were the Hyperkvasir[1], Gastrolab[2] and Gastrointestinal Atlas[3]. The final dataset had 5391 images with approximately 900 images per class. All the images retrieved from the different datasets were labeled by experienced gastrointestinal endoscopists.

4 Implementation of classification models

The chosen CNN architectures to build the classification models were the DenseNet169 (DN169), IRV2, NasNet Large (NNL) and the Resnet50 (RN50). These architectures were selected specific criteria such as what was considered as a good choice performance wise in classification tasks in state-of-the-art studies and the architecture's innovative character.

The models were trained following a transfer learning strategy, using the ImageNet dataset to perform the first training stage. From the pre-trained models, only the feature extractor was kept. The classifier was discarded, given that new fully connected layers were added to the models to build the new classifier. The new layers added to the model were fully connected Dense layers with 64, 32 and 6 nodes. Batch normalisation and dropout layers were also used.

All the models were evaluated on the proposed dataset using 5-Fold Cross-Validation, where 1 subset was used for testing, 1 for validation and the remaining 3 for training. The model parameters such as the optimiser and the learning rate were the algorithm Root Mean Squared Propagation (RMSprop) and 0.001, respectively. Additionaly, no early stopping was performed when training the models and the batch size defined was 32.

The performance metrics chosen to evaluate the models were Accuracy (AC), F1-Score (F1S), Precision (PR), Recall (RC). After each iteration of the Cross-Validation process the results were analysed and, in the end, an average of each metric from the 5 iterations was calculated.

5 Results

Overall, the DN169 was the best model in both classification tasks. The model that came closer to the results obtained by the DN169 was the IRV2, having achieved in the same iterations 0.87 and 0.76 for AC, respectively. The DN169 is the most consistent model and achieves very positive performance results possibly due to the improved conditions to perform the training stage, due to its flow of information and gradients throughout the network. Furthermore, one other important detail is the regularising effect that derives from the dense connections, which reduces the overfitting in tasks with smaller training subsets. Ultimately, the average values for the accuracy performance metric can be seen in Tables 1 and 2.

Table 1: Performance metrics results for Macro classification problems.

CNN model	Macro approach (3-class)			
	AC	PR	RC	F1S
DenseNet169	0.79	0.81	0.79	0.79
Inception-ResnetV2	0.78	0.81	0.78	0.78
NasNet Large	0.72	0.75	0.73	0.73
Resnet50	0.72	0.75	0.73	0.73

Regarding the Grad CAMs, the DN169 was the model used and it focused on a specific part of the analysed images, the region where the lesions, and consequently its features, were located, as seen in Figure 2. Hence, the behaviour of this model was the correct one regarding its performance in the Grad CAMs.

[1] https://osf.io/mh9sj/

[2] https://www.sciencephoto.com/contributor/gas+h9b

[3] https://www.gastrointestinalatlas.com/english/english.html

Table 2: Performance metrics results for Micro classification problems

CNN model	Micro approach (6-class)			
	AC	PR	RC	F1S
DenseNet169	0.72	0.73	0.73	0.71
Inception-ResnetV2	0.65	0.65	0.64	0.63
NasNet Large	0.60	0.60	0.60	0.58
Resnet50	0.57	0.56	0.56	0.54

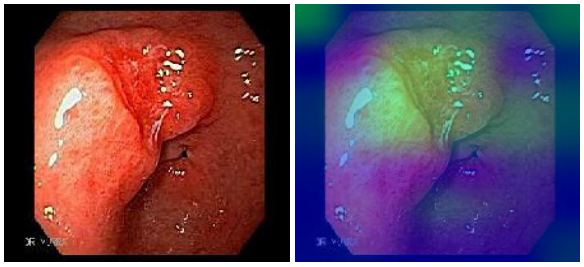


Figure 2: Original AGC image (left) and DN169 Grad CAM image.

6 Conclusions

The present study aimed to solve two different classification tasks by using deep learning algorithms. The subject to be analysed was medical imaging, specifically images from upper endoscopy exams, containing lesions of the Gastrointestinal Tract (GIT). The two classification problems to be solved were of different complexity, one was a multi-class classification problem involving 3 classes and the second one involving 6 classes. The different CNN models were implemented by resourcing to the concept of transfer learning. The models were able to distinguish images without lesions from the ones with a lesion. There was also a trend where the models achieved positive results in separating mild lesions from severe lesions, specifically in the Macro Classes. On the other hand, the models struggled with images which lesions resembled other lesions from a different class. In terms of network architectures, the ones that achieved better results in terms of averaged performance metrics were the DN169 and the IRV2. The DN169 is better prepared when it comes to dealing with overfitting phenomena.

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

References

[1] Freddie Bray et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *A cancer journal for clinicians*, 68:394–424, November 2018. doi: 10.3322/caac.21492.

[2] Bum-Joo Cho et al. Automated classification of gastric neoplasms in endoscopic images using a convolutional neural network. *Endoscopy*, 51:1121–1129, December 2019. doi: 10.1055/a-0981-6133.

[3] Q He et al. Deep learning-based anatomical site classification for upper gastrointestinal endoscopy. *International Journal of Computer Assisted Radiology and Surgery*, 15:1085—-1094, May 2020. doi: https://doi.org/10.1007/s11548-020-02148-5.

[4] Takumi Itoh et al. Deep learning analyzes helicobacter pylori infection by upper gastrointestinal endoscopy images. *Endoscopy International Open*, pages 139–144, February 2018. doi: 10.1055/s-0043-120830.

Improving spatial resolution of myocardial T1-mapping using a model-based super-resolution reconstruction

Francisco Cachado
francisco.cachado@tecnico.ulisboa.pt
Andreia S. Gaspar
andreia.gaspar@tecnico.ulisboa.pt
Rita G. Nunes
ritagnunes@tecnico.ulisboa.pt

ISR-Lisboa/LARSyS and Department of Bioengineering
Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal

Abstract

Over the last years, T_1 mapping has become an important tool for myocardial tissue characterization. For a detailed evaluation, High Resolution (HR) in-plane and sufficient Signal-to-Noise Ratio (SNR) are required, thus, thick slices are often used, sacrificing the through-plane resolution. To address this issue, in this work, a T_1 signal recovery mono-exponential model was combined with a Super-resolution (SR) reconstruction, using the Alternating Direction Method of Multipliers (ADMM) framework as the basis for the iteration system. The algorithm was tested using numerical models; using the proposed approach, it was possible to estimate T_1 values closer to the Ground Truth (GT) maps and the improvement in resolution was noticeable compared to the input T_1 weighted (T_1w) Low Resolution (LR) images.

1 Introduction

Magnetic Resonance Imaging (MRI) is a non-invasive medical imaging technique used for disease diagnosis or treatment monitoring. MRI makes use of nuclear magnetic resonance principles, where magnetic fields and radio frequency pulses are used to excite and then detect changes in the protons' axis to generate images of the tissues in the body. [8]

When setting up an MRI exam, a trade-off between spatial resolution and achieved SNR must be considered. In the clinic it is frequent to opt for high in-plane spatial resolution, sacrificing the through-plane resolution to obtain enough SNR. To increase the through-slice spatial resolution in cardiac MRI (cMRI) parametric mapping, we propose combining a SR reconstruction with a model-based constraint.

1.1 Super-resolution

SR methods aim to increase image resolution. It is a growing area in the Computer Vision field, as it is extremely important to have an image with high pixel density [6]. In this work, SR is used to enable estimation of a HR cMRI T1 map from a set of LR overlapping images.

1.2 T_1 Mapping

T_1 mapping provides a quantitative marker to characterize myocardial tissue without the need for applying contrast agents. This parameter reflects the time required for the longitudinal magnetization, perturbed by the application of radio frequency pulses, to return to equilibrium following a mono-exponential recovery curve.

According to [3, 7], several acquisition protocols and techniques have been developed to perform T_1 mapping; the MOLLI sequence is the gold standard for myocardial T_1 *in vivo* [5] as it requires only one single breathhold. The pulse sequence scheme used in this work is known as the 5(3)3 MOLLI protocol. It consists of 3 phases: the first inversion, where a set of 5 readouts are acquired in consecutive heartbeats; the recovery period, for 3 heartbeats to allow longitudinal magnetization recovery and a second inversion followed by 3 more readouts. Using MOLLI, 8 images are acquired with varying levels of T_1-weighting due to the different inversion times (TI) after application of the inversion pulses. To obtain a T_1-map, a pixelwise fit is carried out as described below.

2 Methods

2.1 Problem formulation

The implemented SR reconstruction method consists in the reconstruction of a HR T_1 map from a set of T_1-weighted (T_1w) LR images.

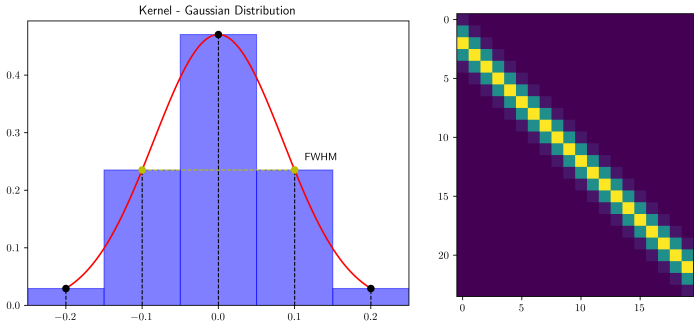
The reconstruction was made considering both data and model consistency terms, resulting in the problem formulated in Eq. 1.

$$\underset{x,u}{\operatorname{argmin}} \|Ax-y\|_2^2 + \lambda \|x-u\|_2 \tag{1}$$

Where x is the reconstructed set of T_1w images and y the set of T_1w LR images. The λ parameter in the second term is the regulatory variable responsible to ensure that the reconstruction is consistent with the model. That is, the images in x for each TI are in accordance with the images u obtained using the signal model. The value of the λ is calculated by using the Eq. 2 and then consecutively divided by 10 until its value is between 1 and 10.

$$\lambda = \left\| A^T \cdot y \right\|_\infty \tag{2}$$

The matrix $\mathbf{A}$ is the convolution matrix, which accounts for the slice profile and shifts in the slice direction [2]. This matrix was built using a specific type of matrix known as Toeplitz matrix, to perform a convolution only relying on the dot operator. The values of each row are obtained through a 1D Gaussian kernel, with a length defined by the user, Fig. 1.



(A) (B)

Figure 1: 1D Gaussian Kernel weights with length 5. (A) Kernel representation in a bar plot. (B) Graphical view of Matrix A with kernel span by the columns.

To tackle the problem formulated in Eq. 1 and since it is a convex optimization problem, the ADMM framework was used [1]. The general ADMM form is:

$$\begin{aligned} &\text{minimize} && f(x)+g(z) \\ &\text{subject to} && Ax+Bz=c \end{aligned} \tag{3}$$

And then the consecutive iterations consists in:

$$x^{k+1} := \underset{x}{\operatorname{argmin}} L_\rho(x, z^k, y^k) \tag{4}$$

$$z^{k+1} := \underset{z}{\operatorname{argmin}} L_\rho(x^{k+1}, z, y^k) \tag{5}$$

$$y^{k+1} := y^k + \rho(Ax^{k+1} + Bz^{k+1} - c) \tag{6}$$

Where the $\mathbf{L}_\rho$ operator is the Lagrangian operator and ρ is the parameter that regulates the influence of the dual variable in the resolution of the global problem.

Taking into account Eq. 3, the minimization problem in Eq. 1 had to be adapted in order to apply the ADMM framework:

$$\begin{aligned} &\text{minimize} && f(x)+g(z) \\ &\text{subject to} && x-z=0 \end{aligned} \tag{7}$$

$$f(x) = \frac{1}{2}\|Ax-y\|_2^2 \tag{8}$$

$$g(z) = \lambda \|z\|_2 \tag{9}$$

Considering the new formulation of the minimization problem, it is necessary to define the iterations:

$$x^{k+1} := (A^T A + \rho I)^{-1}(A^T y + \rho(z^k - u^k)) \quad (10)$$

$$\hat{x} := \mathrm{FIT}(x^{k+1}) \quad (11)$$

$$z^{k+1} := S_{\frac{\lambda}{\rho}}(\hat{x} + u_k) \quad (12)$$

$$u^{k+1} := u^k + \hat{x} - z^{k+1} \quad (13)$$

The fit is done by applying a non-linear least squares algorithm to the mono-exponential model defined in Eq. 14 [4]:

$$M = a - b \cdot e^{\frac{-\mathrm{TI}}{T_1^*}} \quad (14)$$

Where **M** is the T_1w data and **a**, **b** and $\mathbf{T_1^*}$ are the parameters to fit. After the fitting, it is necessary to perform a correction to obtain the true T_1 value, due to MOLLI's readout-induced attenuation of the relaxation curve:

$$T_1 = T_1^*(\frac{b}{a} - 1) \quad (15)$$

2.2 MRXCAT phantom

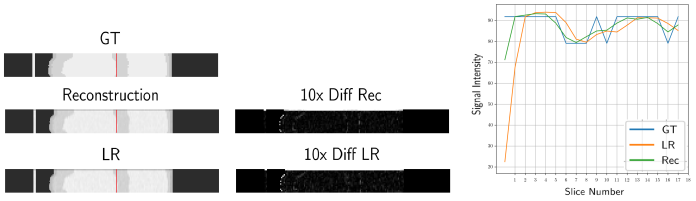
The MRXCAT [9], Fig. 2, is a phantom used for numerical simulation of cMRI, focused on the anatomy of the heart and the surrounding tissues, as well as the movement associated with respiratory and cardiac action.

Figure 2: Atlas from a slice of MRXCAT phantom with labels specifying each tissue type, as provided by its authors [9].

3 Results and Discussions

The Ground Truth (GT) and the Reconstructed (REC) set of the MRXCAT phantom used have a resolution of 1.0×1.0×1.0 mm^3, whereas the LR have a double slice width resulting in a resolution of 1.0×1.0×2.0 mm^3 - these resolutions do not exactly match a typical cMRI protocol; the point was simply to demonstrate recovery of a HR map from LR images. The T_1 map matrix size is 140×180×20. The value used for ρ was 1, and for λ was 7.3, these were set based on a preliminary analysis study where both were allowed to vary and investigating the impact on the reconstruction errors (results not shown).

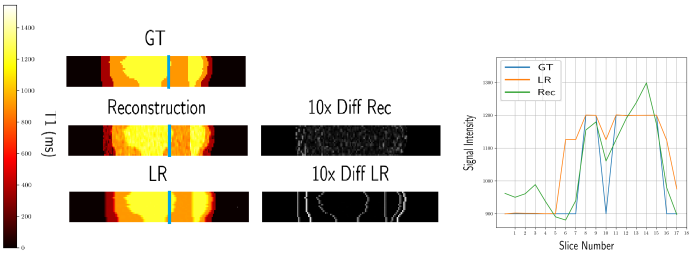
As seen in Fig. 3, the profile recovery in the T_1w images along the z-axis is accomplished with low error values when compared with the GT images, since the Mean Absolute Error (MAE) reduces from 2.84% in the LR to 1.13% in the REC image.



(A) (B)

Figure 3: T_1w image study in the coronal anatomical plane with SNR= 50 and $\rho = 1$. (A) T_1w GT, LR and Reconstructed images and the difference between the last two and GT. (B) Line profile for each set of images, the slice used is marked with red in (A).

Regarding the T_1 map reconstruction, Fig. 4, the reconstruction has also been successful with a MAE decreasing from 2.26% to 1.58%. Nevertheless, it is important to notice that in the interface regions the reconstruction shows some differences when compared with the GT, as visible in the line profile in Fig. 4B.



(A) (B)

Figure 4: T_1 map study in the sagittal anatomical plane with SNR= 50 and $\rho = 1$. (A) T_1 GT, LR and REC Maps and the difference between the last two and GT. (B) Line profile for each set of maps, shown in green in the used slice (A).

The reconstruction's convergence can be assessed by analysing both residuals' evolution, and as seen in Fig. 5, both converged, although, the primal residual fluctuates slightly after it had already converged.

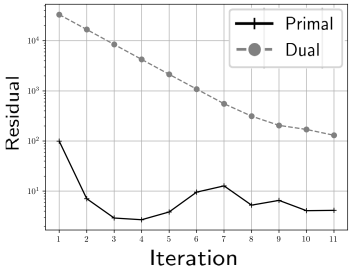


Figure 5: Primal and Dual residual evolution for the reconstruction of the MRXCAT phantom with SNR= 50 and $\rho = 1$ during 11 iterations.

4 Conclusion

The resolution was successfully improved with the proposed method. Regarding the results presented, the profile recovery along the z-axis are in line with an accurate application of a SR reconstruction. The recovery of the profile in T_1w is very good, whereas in the case of the T_1 map it has a slightly higher error. Nonetheless, both reconstructed signals show very low error values when compared to GT, around 1%.

With this work, it was possible to demonstrate the application and importance of SR in the Medical Imaging sector. To achieve the goal of providing medical professionals with cMRI HR images in a real-life situation, *in vivo* validation is necessary. For that, information regarding the positional slice shifts associated to residual respiratory movement would need to be included in the reconstruction.

Acknowledgements

FCT (SFRH/BD/120006/2016, PTDC/EMD-EMD/29686/2017, UIDP/50009/2020); Lisboa 2020 (LISBOA-01-0145-FEDER-029686)

References

[1] Boyd S. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2010.

[2] Hilbert T., et al. Fast model-based t 2 mapping using SAR-reduced simultaneous multislice excitation. *Magnetic Resonance in Medicine*, 82(6):2090–2103, July 2019.

[3] Kim P. K., et al. Myocardial t1 and t2 mapping: Techniques and clinical applications. *Korean Journal of Radiology*, 18(1):113, January 2017.

[4] Kim Y.-C., et al. Fast calculation software for modified look-locker inversion recovery (MOLLI) t1 mapping. *BMC Medical Imaging*, 21(1), February 2021.

[5] Messroghli D. R., et al. Optimization and validation of a fully-integrated pulse sequence for modified look-locker inversion-recovery (MOLLI) t1 mapping of the heart. *Journal of Magnetic Resonance Imaging*, 26(4):1081–1086, October 2007.

[6] Moran M. B. H., et al. Using super-resolution generative adversarial network models and transfer learning to obtain high resolution digital periapical radiographs. *Computers in Biology and Medicine*, 129:104139, February 2021.

[7] Roujol S., et al. Accuracy, precision, and reproducibility of four t1 mapping sequences: A head-to-head comparison of MOLLI, ShMOLLI, SASHA, and SAPPHIRE. *Radiology*, 272(3):683–689, September 2014.

[8] Smith N. B. et al. Magnetic resonance imaging (mri). In *Introduction to Medical Imaging: Physics, Engineering and Clinical Applications*, pages 204–282. Cambridge University Press, Cambridge, 2009.

[9] Wissmann L., et al. MRXCAT: Realistic numerical phantoms for cardiovascular magnetic resonance. *Journal of Cardiovascular Magnetic Resonance*, 16(1):63, August 2014.

Deep Convolutional Neural Network for gastric landmark detection

Inês Lopes
ines.videiralopes@ua.pt
INESC TEC, Physics Department,
Aveiro University

Miguel Coimbra
mcoimbra@dcc.fc.up.pt
INESC TEC,
Faculty of Sciences, University of Porto

Augusto Silva
augusto.silva@ua.pt
Department of Electronics, Telecommunications, and Informatics
Aveiro University

Francesco Renna
frarenna@dcc.fc.up.pt
Instituto de Telecomunicações, INESC TEC,
Faculty of Sciences, University of Porto

Abstract

Gastric cancer is the fifth most incident cancer in the world and, when diagnosed at an advanced stage, its survival rate is only 5%-25%, providing that it is essential that the cancer is detected at an early stage. However, physicians specialized in this diagnosis have difficulties in detecting early lesions during a diagnostic examination, esophagogastroduodenoscopy (EGD). Early lesions on the walls of the digestive system are imperceptible and confounded with the stomach mucosa, being difficult to detect. On the other hand, physicians run the risk of not covering all areas of the stomach during diagnosis, especially areas that may have lesions. The introduction of deep learning (DL) into this diagnostic method may help to detect gastric cancer at an earlier stage. The aim of this work consists in testing new automatic algorithms, specifically CNN-based systems able to detect upper gastrointestinal (GI) landmarks to avoid the presence of blind spots during EGD to increase the quality of endoscopic exams. We tested some pre-trained architectures as the ResNet-50, DenseNet-121, and VGG-16. For each pre-trained architecture, we tested different learning approaches, including the use of class weights (CW), the use of batch normalization (Bn) and dropout layers, and the use of data augmentation (DA) to train the network. The CW ResNet-50 achieves an accuracy of 71.79% and a Mathews Correlation Coefficient (MCC) of 65.06%. In addition, we tested CW ResNet-50 concatenation with convolutional autoencoder models, and we achieve an accuracy of 72.14% and an MCC of 64.88%.

Keywords *- Convolutional Neural Network, Autoencoder Network, Upper Gastrointestinal Landmarks, Esophagogastroduodenoscopy*

1 Introduction

There are several studies that prove the ability of CNNs in the monitorization of blind spots. However, the CNNs investigations in the detection of gastric landmarks are still in the beginning.

The *Kvasir* dataset is a collection of GI images that became available in 2017, it promoted CNNs exploration in landmark detection, with several studies showing promising results. Pogorelov et al. [1] uses architectures with 3 and 6 convolutional layers, which presented good performances. However, Pogorelov et al. concluded that using a pre-trained architecture like the Inception-v3 gave better performance. Agrawal et al. [2] used CNN architectures (VGG-16 and Inception-v3) to extract features and obtained accuracies above 0.95. These results show the ability of CNNs to extract the features that better represent the images. Petscharning et al. [3] used an inception-like CNN architecture focusing on variables, such as the number of neurons in the network and the size of the training repository. Petscharning et al. concluded that better performances are obtained with large training datasets and many parameters could compromise the network performance, generating overfitting. It is important to maintain a balance between the number of parameters and the size of the training repository. Cogan et al. [4] tested several architectures pre-trained over ImageNet to classify the *Kavasir* repository. In this study, Inception-v4, Inception-ResNet-v2, and NASNet networks were tested, which all had accuracies above 0.97. However, the NASNet performance is noteworthy because it is an architecture that uses CNNs with RNN controllers recursively.

The Borgli et al. [5] used the *HyperKvasir* repository that has 3 upper GI anatomical landmarks and tested several architectures (ResNet and DenseNet models) as a baseline for future studies and to show the dataset potential.

All previous studies only classified 2/3 upper GI anatomical landmarks. However, it is important to evaluate the CNNs performances with more anatomical landmarks. Takiyama et al. [6] used a superior number of EGD images and anatomical landmarks (collected from a hospital) than the *Kvasir* repository and used a GoogLeNet architecture, which showed an area under the curves (AUCs) above 0.99. He et al. [7] used several CNN architectures pre-trained with ImageNet repository (ResNet-50, Inception-v3, VGG-11-bn, VGG-16-bn, DenseNet-121) and used 11 anatomical landmarks, requiring greater efficiency by CNN architectures. However, all CNN architectures had accuracies above 0.87, which means that, with an increase of anatomical landmarks, the CNNs continue to perform well. Wu et al. [8] increased the anatomical landmarks to 26 and compared the performance with only 10 anatomical landmarks with the same architecture (VGG-16). It is noticeable that the accuracy decreases greatly with the increase of anatomical landmarks: the accuracy with 10 landmarks was 0.9, while with 26 landmarks it went to 0.659. The complexity of the architecture increases with more landmarks.

This work provides the following novel contributions: **1.** The test of different pre-trained CNN architectures on a new anatomical landmark's dataset assembled with images from public repositories; **2.** The study of the impact of the use of autoencoders to extract meaningful features from EGD images leveraging a large amount of unlabelled data.

2 Material and Methods

2.1Data

Firstly, we started by searching datasets that contain upper GI tract images and videos with healthy anatomical sites and with pathologies. The repositories that were used are *HyperKvasir* [5] and GASTROLAB [9]. One frame per second (fps) was extracted from GI videos. We obtained 9 classes: Esophagus with 2 176 frames; Z-line with 1 679 frames; Fundus with 135 frames; Cardia with 652 frames; Retroflex Stomach with 838 frames; Body with 988 frames; Antrum and Pylorus with 2 124 frames; Duodenal Bulb with 417 frames; Duodenum with 1 828 frames.

2.2Architectures

The used pre-trained architectures were ResNet-50, DenseNet-121, and VGG-16:

- **Pt ResNet-50 experiment** is based on pre-trained ResNet-50 architecture by *HyperKvasir* baseline [5], where we adjusted the architecture to our task of classifying 9 different upper GI classes. We took the layers from pre-trained ResNet-50 with ImageNet weights and froze them to avoid destroying any information. Then, we added a global average pooling layer and a dense layer on top of the frozen layers to predict the respective classes.
- **CW ResNet-50 experiment** has the same architecture as the Pt ResNet-50, but in this case, class weights are used in the definition of the loss function. This is a strategy to try to attenuate the unbalanced dataset effect.
- **DA ResNet-50 experiment** has the same architecture as Pt ResNet-50, but we used data augmentation for training. We applied 18 transformations to each class to balance the dataset. We applied Gaussian noise, rotations, flipping, transposing, cropping, adjusting saturation, and adjusting brightness.
- **Bn ResNet-50 experiment** has a base model equal to the Pt ResNet-50. However, we added a batch normalization layer. We also swapped the global average pooling layer for the global max pooling layer since max pooling is a noise suppressant. We added two more dense layers than Pt ResNet-50 and we added a dropout layer to the model, whose goal is to reduce overfitting.

We repeated the previous four experiments (class weights, batch normalization, and data augmentation) for pre-trained DenseNet-121 and pre-trained VGG-16.

Autoencoder architectures were also tested in our task. The autoencoder models were trained with 99 417 unlabelled images from the *HyperKvasir* repository. These architectures' main feature consists of input image reconstruction through the extraction of features with

convolutional layers. The autoencoder architecture consists of two parts: an encoder and a decoder. In the next paragraph, the experiments with autoencoder architectures are described:

- **Convolutional encoder (CE) experiment** consists of using the encoder layers from the autoencoder model built with unlabelled images, as shown in Figure 1. We took the encoder layers trained with unlabelled images and froze them to avoid destroying any information. Then, we added the global average pooling layer and a dense layer on top of the frozen layers to predict the respective classes. We also use class weights for training.

- Concatenation of CE with CW ResNet-50 (**SCE + CW ResNet-50 experiment**) consists of concatenating the encoder section (CE) with the CW ResNet-50 model (see Figure 2). We added a global average pooling layer on the top of each concatenation section (SCE and CW ResNet-50). Then, we added a dense layer on top of the frozen layers to predict the respective classes (see Figure 2).

3 Experiments

To validate our models, we applied 5-fold cross-validation and we evaluated their performance in terms of macro-average, and weighted-average precision, recall, and F1-score. Additionally, we calculated the accuracy and the MCC.

The overall results are reported in Table 1. The model that presents the best performance is the CW ResNet-50 with a MCC of 65.06% (see Table 1). The performance values of the Pt ResNet-50 architecture are very close to the CW ResNet-50 architecture. However, the Pt ResNet-50 architecture improved its performance with the addition class weights. In contrast, the use of data augmentation in the training of Pt ResNet-50 architecture did not improve the performance, which have a MCC of 60.90%. Contrary to what we expected, the Bn ResNet-50 architecture has the worst performance, the MCC was only 37.03%. So, the CW ResNet-50 architecture continues to have the best performance.

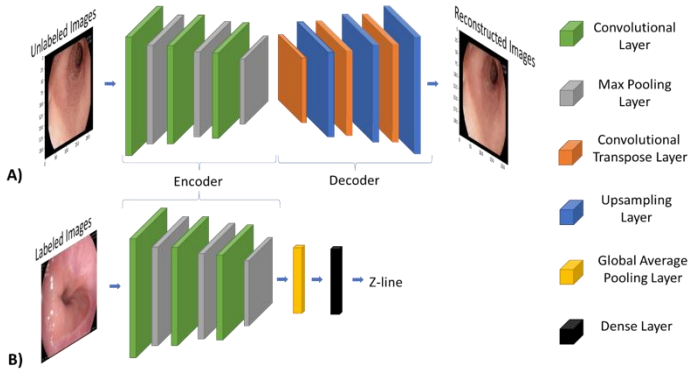


Figure 1: A) The convolutional autoencoder architecture, which was trained according to the unlabelled GI images from *HyperKvasir* repository. B) The CE model, which corresponds to the encoder part of A). The CE was trained to classify labelled anatomical zones. We froze encoder layers to avoid destroying any information learning in A).

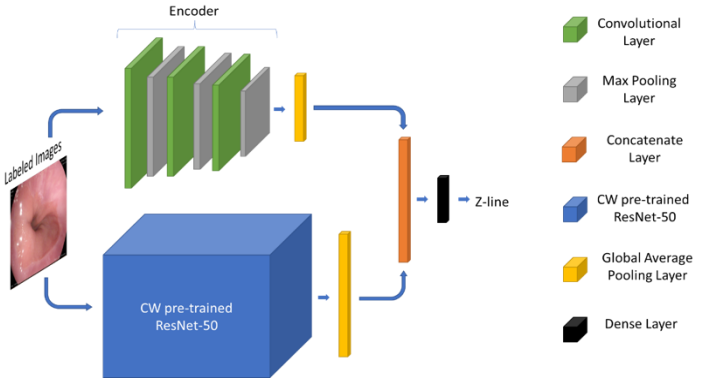


Figure 2: SCE + CW ResNet-50 experiment.

Therefore, the ResNet-50 architectures are the ones that best result in the classification of anatomical zones. VGG-16 architectures perform better than DenseNet-121 architectures.

CE experiment does not present good performance (MCC of 32.85%), however, their performances improve considerably with CW ResNet-50 concatenation (MCC of 64.88%). In general, the concatenation of features extracted with a convolutional autoencoder does not significantly improve the classification of anatomical zones. Comparing the metrics values of CE + CW ResNet-50 experiment with CW ResNet-50 experiment, it is visible that the values are close (see Table 1).

Table 1: Performance metrics of the pre-trained ResNet-50 experiments.

Models	Macro-average			Weighted-average			Accuracy	MCC
	Precision	Recall	F1-score	Precision	Recall	F1-score		
Pt ResNet-50	63.22%	59.89 %	59.91 %	69.55%	69.53 %	68.56 %	72.40%	64.35 %
DA ResNet-50	57.75%	57.80 %	56.20 %	65.57%	66.47 %	64.91 %	69.56%	60.90 %
CW ResNet-50	64.10%	61.50 %	61.50 %	70.65%	70.01 %	69.26 %	71.79%	65.06 %
Bn ResNet-50	32.54%	33.49 %	31.18 %	43.15%	47.61 %	43.00 %	49.78%	37.03 %
CE	30.63%	30.29 %	27.09 %	39.27%	32.04 %	32.45 %	45.45%	32.85 %
CE + CW ResNet-50	62.85%	61.32 %	60.58 %	71.13%	69.77 %	69.11 %	72.14%	64.88 %

4 Conclusions

In this work, we have considered the use of different CNN architectures to classify anatomical landmarks from EGD images. Different techniques have been applied in the attempt to reduce overfitting, including transfer-learning, data augmentation, and the use of latent space representations learned from a large dataset of unlabeled images. In state-of-art studies, only supervised learning approaches were used to classify EGD images, and in our work, we used unsupervised learning to help our task. However, the use of features learned in an unsupervised fashion via the application of autoencoder did not provide a significant boost in the classification performance.

Acknowledgments

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

References

[1] K. Pogorelov *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," *Proc. 8th ACM Multimed. Syst. Conf. MMSys 2017*, pp. 164–169, 2017, doi: 10.1145/3083187.3083212.

[2] T. Agrawa, R. Gupta, S. Sahu, and C. E. Wilson, "SCL-UMD at the medico task-mediaeval 2017: Transfer learning based classification of medical images," *CEUR Workshop Proc.*, vol. 1984, pp. 3–5, 2017.

[3] S. Petscharnig, K. Schoffmann, and M. Lux, "An inception-like CNN architecture for GI disease and anatomical landmark classification," *CEUR Workshop Proc.*, vol. 1984, pp. 0–2, 2017.

[4] T. Cogan, M. Cogan, and L. Tamil, "MAPGI: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning," *Comput. Biol. Med.*, vol. 111, no. April, p. 103351, 2019, doi: 10.1016/j.compbiomed.2019.103351.

[5] H. Borgli *et al.*, "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Sci. Data*, vol. 7, no. 1, p. 283, Dec. 2020, doi: 10.1038/s41597-020-00622-y.

[6] H. Takiyama *et al.*, "Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks," *Sci. Rep.*, vol. 8, no. 1, pp. 1–8, 2018, doi: 10.1038/s41598-018-25842-6.

[7] Q. He *et al.*, "Deep learning-based anatomical site classification for upper gastrointestinal endoscopy," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 7, pp. 1085–1094, 2020, doi: 10.1007/s11548-020-02148-5.

[8] L. Wu *et al.*, "Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy," *Gut*, vol. 68, no. 12, pp. 2161–2169, 2019, doi: 10.1136/gutjnl-2018-317366.

[9] "GASTROLAB - the Gastrointestinal Image Site." www.gastrolab.net.

Segmentation of US fetus images based on particle swarm optimization and k-means clustering

Lio Gonçalves
lgoncalv@utad.pt
UTAD-ECT-Departamento de Engenharias
The INESC-TEC-Institute for Systems and Computer Engineering, Technology and Science
Porto, Portugal

Paulo Salgado
psal@utad.pt
ECT-Departamento de Engenharias
Universidade de Trás-os-Montes e Alto Douro
Vila Real, Portugal

Paulo Afonso
pafnaa@ua.pt
Escola Superior de Tecnologia e Gestão de Águeda
Universidade de Aveiro,
Aveiro, Portugal

Abstract

Fetal ultrasound image segmentation is a topic with new advancements that allow doctors to diagnose fetal structural abnormalities such as those involved in gestational diabetes mellitus, pulmonary sequestration, congenital heart disease, etc. The new technologies provide more insight about the development of the fetus. The image segmentation of the whole fetus, in particular, brain, lungs, heart, liver etc is vital for clinicians to get more knowledge of the anatomy of the fetus without miss diagnose as the US images are very "noisy". In this work we propose a new method to do the image segmentation of the amniotic sac. The algorithm is inspired in the well known Particle Swarm Optimization algorithm (PSO). The proposed algorithm combines PSO with k-means clustering and several operations, as crossover and mutation, among others, to improve the convergence towards the region that corresponds to the amniotic sac. Particles of this so-called PSO-K-C-M algorithm are trajectories obtained from the interpolation of periodic cubic splines.

1 Introduction

Medical ultrasonography is the use of medical ultrasonic equipment's and imaging techniques to visualize internal organs to capture their size, structure and any pathological lesions with real time tomographic images. Obstetric ultrasonography is the use of medical ultrasonography in pregnancy to create real-time visual images of the developing embryo in its mother's uterus (womb). Image segmentation of US fetus images is an hard task as images have poor quality,e.g., low contrast and high level of noise.

The reader is invited to read the approaches on US image segmentation proposed by Zong *et al*. [5]. For an extensive and well done review about segmentation and classification in MRI and US fetal imaging check the work developed by Torrents-Barena *et al*. [1]. The review covers state-of-the-art segmentation and classification methodologies for the whole fetus and several organs.

In a previous work, presented at RECPAD 2019 [2], the authors proposed a method to tackle the problem of segmentation of the fetus from the original US fetus image. Now we propose an hybrid algorithm that combines PSO standard algorithm, with k-means clustering and several operations as crossover or mutation, to get an "envelope" of the amniotic sac. Particles of this so-called PSO-K-C-M algorithm are trajectories that derive from the interpolation of periodic cubic splines.

2 Particle Swarm Optimization and K-means Clustering

2.1 Particle Swarm Optimization (PSO)

Particle Swarm Optimizton is a method developed for finding a global optima of some nonlinear function [3]. It has been inspired by a social behavior of birds and fish. Each solution consists of set of parameters and represents a point in multidimensional space. The solution is called *particle* and the group of particles (population) is called *swarm*. These particles are moved around in the search-space according to a few simple formulae, they are guided by their own best known position in the search-space as well as the entire swarm's best known position, iteratively trying to improve a candidate solution with regard to a given measure of quality.

The process is described as follows: Each particle i is represented as a D-dimensional position vector $X_i = (x_{i1}, x_{i2}, \ldots x_{id})$ and has a corresponding instantaneous velocity vector $V_i = (v_{i1}, v_{i2}, \ldots v_{id})$. The best position experienced by particle during the process of update and iteration can be expressed as the D-dimensional vector $P_i = (p_{i1}, p_{i2}, \ldots p_{id})$ and the best position of the entire swarm as vector $P_s = (p_{s1}, p_{s2}, \ldots p_{sd})$.

The update rule for the velocity of the particle i is given by:

$$v_{ij}(k+1) = w\, v_{ij}(k) + c_1 r_1 \left[p_{ij} - x_{ij}(k)\right] + c_2 r_2 \left[p_{si} - x_{ij}(k)\right], \quad (1)$$

where w is the inertia coefficient, c_1 and c_2 are learning factors,respectively representing self-learning capacity and capacity to learn "from swarm" behaviour; r_1 and r_2 are two random numbers of average distribution of interval $[0,1]$.

The update rule for the position of the particle i is shown as follows:

$$x_{ij}(k+1) = x_{ij}(k) + v_{ij}(k+1),\ j = 1, \ldots, d \quad (2)$$

2.2 K-means Clustering

K-means is an unsupervised learning algorithm that solves the well-known clustering problem because of its fast execution and easy implementation [4].

To classify a given data set into a fixed number of clusters (assume k clusters), it defines k centers, one for each cluster. These centers should be placed far away from each other for better chances to get global optimal solution. Then associate each data point to any of these clusters having nearest center. Then re-calculate k new centers as bary centers of the clusters and rebind the same data set points to nearest new center. This discrete movement of the centers can be understood as result of the action of "forces" applied to the centers. This new "force" constitutes an additional part of equation (1).

Repeat this process either for a fixed number of iterations or until two sub-sequent iteration having same centers.

3 PSO-K-C-M algorithm

The PSO-K-C-M proposed in this work stands for an hybrid algorithm that includes, *PSO* algorithm, with *K*-means clustering, *C*rossover, *M*utation, and other variants. The procedure corresponds to the following steps, among others:

Step 1 Image pre-processing, e.g., filtering, binarization, negative image, thresholding.

Step 2 Initialization of population of "particles" (trajectories through cubic splines interpolation), for PSO optimization.

Step 3 How particles perform? Compute pixel gradients and global "pixel intensity" of particles (trajectories). Record personal best for particles and global best for swarm.

Step 4 Standard PSO algorithm unfolds.

Step 5 Choose a fraction of the PSO population and apply *k-means clustering* "to push" those particles to white pixels region (amniotic sac).

Step 6 Use *crossover* to sew some sections of the trajectories, obtained from the particles with better performance, as if it were a cardiac bypass. Without forgetting memory of personal best.

Step 7 Use *mutation* to change some particles; keep personal best.

Step 8 *Switch* personal best between an fraction of particles.

Step 9 Change the speed of some particles; should be less or equal maximum velocity allowed.

Step 10 Repeat Steps 4-9 while Performance index keeps improving.

4 Results

The following images illustrate some steps of the proposed method, e.g., the original image in Figure 1; The Global Best particle at first iteration, among 50 particles, is depicted in Figure 2. Figure 3 illustrates the Global Best particle of the swarm on last iteration. At Figure 4, 50 trajectories/particles are depicted-at last iteration. The evolution of the Performance index and the particles enhanced with the incorporation on PSO algorithm of K-means clustering/Crossover/Mutation are depicted in the last figures.

Figure 1: Original image.

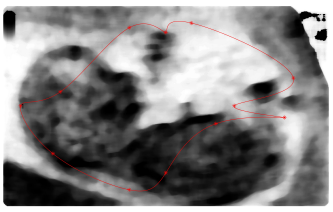


Figure 2: Global Best particle- 1st iteration.

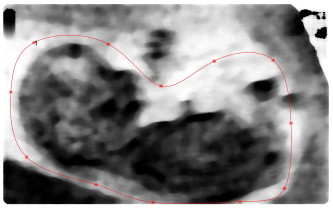


Figure 3: Global Best particle- last iteration.

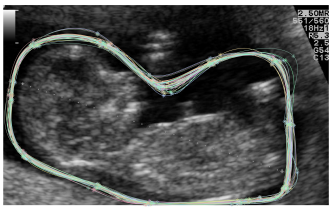


Figure 4: 50 particles/trajectories-last iteration.

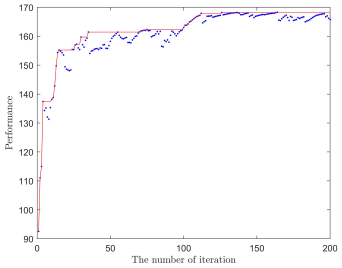


Figure 5: Evolution of Performance index.

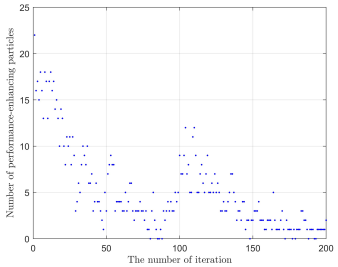


Figure 6: Enhanced particles (crossover/mutation/etc) by iteration.

5 Conclusions

The aim was to develop an algorithm that would allow segmentation of the amniotic sac in US images of fetuses. The approach taken was to use trajectories - which are obtained from the interpolation of cubic (periodic) splines - in order to obtain an "envelope" of the amniotic sac (at least, the largest possible area). The proposed PSO-K-C-M algorithm consists of an hybrid procedure that is obtained from the standard PSO algorithm combined with some variants, such as k-means clustering, crossover and mutation, in order to obtain better performance. The results were very promising. Note that unlike emerging image segmentation techniques, such as applications of machine learning or deep learning, this approach does not require a vast database to achieve good results.

References

[1] J. T. Barrena, G. Piella, N. Masoller, E. Gratacós, E. Eixarch, M. Ceresa, and M. Ballester. Segmentation and classification in mri and us fetal imaging: Recent trends and future prospects. *Medical Image Analysis*, 51:61 – 88, 2019.

[2] L. Gonçalves, P. Salgado, and P. Afonso. Segmentation of us fetus images through dijkstra's inspired algorithm. In *25th Portuguese Conference on Pattern Recognition*, 2019.

[3] James Kennedy. *Particle Swarm Optimization*, pages 760–766. Springer US, Boston, MA, 2010.

[4] K. R. Zalik. An effcient k'means clustering algorithm. *Pattern Recognition Letters*, 29:1385 – 1391, 2008.

[5] J. J. Zong, T. S. Diu, W. D. Li, and D. M. Guo. Automatic ultrasound image segmentation based on local entropy and active contour model. *Computers Mathematics with Applications*, 78(3):929 – 943, 2019.

Anonymising Case-based Explanations for Medical Image Analysis

Helena Montenegro[1,2]
up201604184@edu.fe.up.pt
Wilson Silva[1,2]
wilson.j.silva@inesctec.pt
Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt

[1] Faculdade de Engenharia
Universidade do Porto
Porto, Portugal
[2] INESC TEC
Porto, Portugal

Abstract

Case-based interpretability provides intuitive explanations for Deep Learning models' decisions. In clinical contexts, the use of visual case-based explanations raises privacy concerns, as they expose patient data, requiring anonymisation. In this work, we analyse and compare existing visual anonymisation methods applied to anonymise medical case-based explanations. We conclude that the existing methods are not sufficiently developed, as they do not guarantee the three fundamental requirements of case-based explanations: privacy, explanatory evidence and intelligibility.

1 Introduction

Deep Learning has achieved outstanding results in medical image analysis tasks. However, the "black-box" nature of these models hinders their application in clinical contexts, as their decisions are difficult to understand and trust. Case-based interpretability methodologies provide intuitive visual explanations through the retrieval of diagnostic cases from data. However, these explanations disclose patient data, which must be anonymised to protect the patients' privacy.

Current anonymisation methods for visual data can be divided into traditional and deep learning methods. Traditional methods are applied to the whole input, requiring an additional pre-processing step to identify the image parts that need to be anonymised [2, 3]. Deep learning methods learn to identify and modify the image regions that disclose identity to attain privacy [1]. In this work, we analyse and compare visual anonymisation techniques applied to medical case-based explanations. We evaluate their capacity to preserve the fundamental requirements of the intended explanations: privacy, explanatory evidence and intelligibility.

2 Method

We apply the visual anonymisation methods to a medical and biometric dataset of iris images for glaucoma recognition: Warsaw-BioBase-Disease-Iris v2.1 [6, 7].

As traditional methods, we used blur [2] and K-Same-Select [3], and varied the parameters of each method to investigate the trade-off between privacy, intelligibility and explanatory evidence. In blur, we apply Gaussian kernels of different sizes to the images. As K-Same-Select averages images grouped in K-sized clusters organised by class, we vary K, corresponding to the number of identities per averaged image.

As a deep learning model, we selected the only method from the literature that considers the preservation of task-related features: PPRL-VGAN [1]. The model anonymises through identity replacement, outputting an image where the patient provided as the replacement can be recognised. The architecture of the model, as shown in Figure 1, comprises a Generative Adversarial Network with a Variational Autoencoder as the generator. To aid the generator's training, the network contains a multi-task discriminator with a real/fake classifier to promote realism in the synthetic images, an identity recognition network to guide the identity replacement process, and a task-related classifier to preserve the original image's class.

In the evaluation, we use an identity recognition network and a glaucoma recognition network. To evaluate privacy, we analyse the accuracy of the identity recognition network at recognising the original subject and identities used in the anonymisation process (as a replacement or for averaging). To evaluate explanatory evidence, we verify the accuracy and F1-score of a glaucoma recognition network at identifying the pathology in the anonymised image. Furthermore, we use the saliency map method Deep Taylor [4] to visualise whether the semantic features of the anonymised images are similar to the ones in the original images.

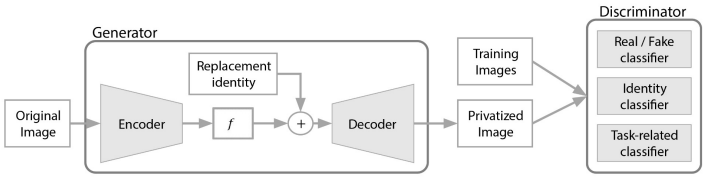


Figure 1: PPRL-VGAN model's architecture [5]

3 Results

The results are summarised in Table 1, with the best results for each method highlighted in bold.

An example of using blur is shown in Figure 2. The accuracy of the identity recognition network is higher than in any other method used, suggesting a poor privacy-preserving capacity. Furthermore, the intelligibility and the explanatory evidence of the anonymised images decrease with higher levels of blurring, hindering the images' use as explanations.

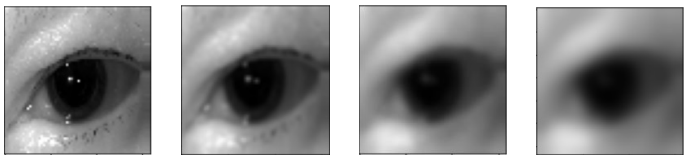


Figure 2: Results of blurring the original image (a) with kernels of dimensions 3, 9 and 15 (b-d).

Some results obtained with K-Same-Select are shown in Figure 3. As K increases, the accuracy in both recognition networks decreases, guaranteeing higher privacy but lower explanatory value. This method fails at preserving explanatory evidence.

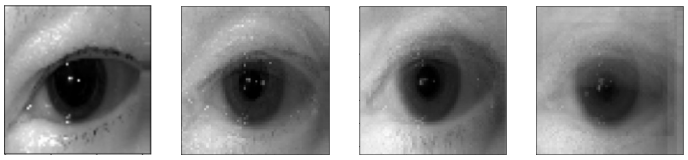


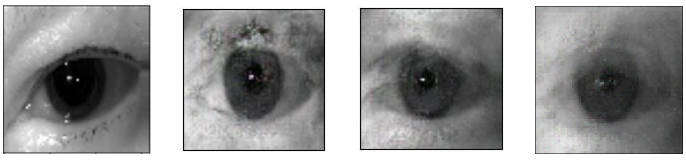
Figure 3: Results of applying K-Same-Select to the original image (a) with 3, 6 and 9 (b-d) as the values of K.

With PPRL-VGAN, we performed various experiments by changing the patients we selected as the replacement identities. We started by selecting randomly among all patients, which resulted in unexpectedly low glaucoma recognition results. Then, we selected the replacement identities from patients that share the pathology of the original image, which significantly improved the glaucoma recognition scores. Finally, we selected the replacement identities among patients whose pathology differs from the original image, leading to significantly worse results in glaucoma recognition. As such, the PPRL-VGAN model struggles at reproducing disease-related features in the anonymised images when using identities that do not possess the pathology of the original image. In terms of

Experiment	Dataset	Identity Recognition Accuracy (↓)	Replacement Identity Recognition Accuracy (↓)	Glaucoma Recognition Accuracy (↑)	Glaucoma Recognition F1 Score (↑)
Baseline	Original test set	90.00%	-	93.24%	87.83%
PPRL-VGAN	Anonymised set w/ random identities	**0.50%**	74.68%	79.18%	62.22%
	Anonymised set w/ identities w/ the same pathologies	1.76%	78.35%	**86.56%**	76.01%
	Anonymised set w/ identities w/ different pathologies	0.71%	60.26%	65.06%	48.30%
	Averaged anonymised set	2.56%	**14.35%**	86.24%	**78.80%**
Blurring	Anonymised set with kernel size 3	69.41%	-	**93.24%**	**87.57%**
	Anonymised set with kernel size 9	31.76%	-	88.82%	77.11%
	Anonymised set with kernel size 15	**23.24%**	-	81.47%	55.32%
K-Same-Select	Anonymised set with 3 identities	7.06%	22.94%	**82.35%**	**61.54%**
	Anonymised set with 6 identities	2.94%	**14.41%**	81.76%	53.73%
	Anonymised set with 9 identities	**1.47%**	**14.41%**	78.53%	42.52%

Table 1: Experiments Results [5].

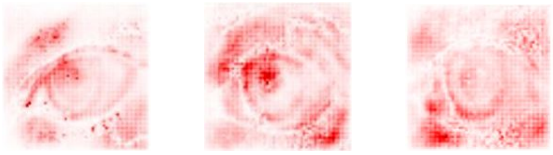
privacy, all the experiments were capable of preserving privacy for the original patient. However, the model discloses the identity used as a replacement, as evidenced by the high accuracy in the replacement identity recognition. To improve this problem, we performed an additional experiment where we averaged 6 images anonymised with PPRL-VGAN using patients with the pathology of the original image as a replacement. This experiment provided the most balanced results, with low identity recognition accuracy and high glaucoma recognition scores. Some examples of anonymised images obtained with this method are shown in Figure 4.



(a) (b) (c) (d)

Figure 4: Results of applying the PPRL-VGAN model to the original image (a) with a replacement identity with the same pathology (b) and with a different pathology (c). (d) is from the averaged anonymised set.

Using Deep Taylor [4] in the anonymised images acquired with PPRL-VGAN, we obtained the results shown in Figure 5. The anonymised image with glaucoma displays pixels with higher relevance for the classification in the same regions as the original image (upper side of the iris). Furthermore, these regions are not highlighted in the anonymised image that does not contain glaucoma (c). As such, the anonymisation process preserves semantic features.



(a) (b) (c)

Figure 5: Results of applying Deep Taylor to glaucoma recognition in the original image with glaucoma (a) and in anonymised images with glaucoma (b) and without glaucoma (c).

4 Discussion and Conclusions

We evaluated different visual anonymisation methods from three perspectives essential to case-based explanations: privacy, preservation of explanatory evidence, and intelligibility.

Blur cannot achieve satisfying levels of privacy while preserving explanatory evidence and image intelligibility. K-Same-Select guarantees privacy to the extent of K-Anonymity, i.e., the highest probability of a patient being recognised is $\frac{1}{K}$ [3]. However, K-Anonymity may not be enough to protect the patients' privacy rights. This method also fails to preserve the explanatory features of the original image.

PPRL-VGAN violates the privacy of patients used as a replacement during anonymisation. The method fails at preserving explanatory evidence when using patients with a different pathology than the original image as a replacement. Nonetheless, the method can preserve explanatory features using patients that share the pathology of the original image. Averaging anonymised images improved the results in terms of privacy. However, it has the same drawbacks as the K-Same-Select method, as its privacy is limited to K-Anonymity, and some relevant semantic features may be lost.

To conclude, the experiments suggest the need to improve anonymisation methods for medical case-based explanations. The privacy-preserving methods should fulfil the following requirements: privacy for all patients in the dataset, the preservation of the exact explanatory features of the original image, and intelligibility. The existence of a privacy-preserving model that guarantees these requirements would enable the use of case-based explanations in clinical contexts to improve trust and acceptance of deep learning and, consequently, the quality of medical diagnosis.

5 Acknowledgements

This work was partially funded by the Project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership, and also by the Portuguese Foundation for Science and Technology - FCT within PhD grant number SFRH/BD/139468/2018.

References

[1] J. Chen, J. Konrad, and P. Ishwar. Vgan-based image representation learning for privacy-preserving facial expression recognition, 2018.

[2] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent. Large-scale privacy protection in google street view. In *ICCV 2009*, pages 2373–2380, 2009.

[3] R. Gross, E. Airoldi, B. Malin, and L. Sweeney. Integrating utility into face de-identification. In *Privacy Enhancing Technologies*, pages 227–242, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[4] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[5] H. Montenegro, W. Silva, and J. S. Cardoso. Towards privacy-preserving explanations in medical image analysis. In *Proceedings of the 1st Workshop on Interpretable Machine Learning in Healthcare (IMLH), as part of the ICML conference*, 2021.

[6] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Assessment of iris recognition reliability for eyes affected by ocular pathologies. In *BTAS Conference 2015*, pages 1–6, 2015.

[7] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Implications of ocular pathologies for iris recognition reliability. *Image and Vision Computing*, 58:158–167, 2017.

Automatic Lung Field Segmentation on Chest Radiography Images

Rui Magalhães[1]
rui.magalhaes@tecnico.ulisboa.pt
Ricardo Brioso[2,3]
Joana Rocha[2,3]
Sofia Cardoso Pereira[2,3]
João Pedrosa[2,3]
Ana Maria Mendonça[2,3]
Aurélio Campilho[2,3]

[1] Instituto Superior Técnico (IST) Lisboa, Portugal
[2] Faculty of Engineering of the University of Porto (FEUP) Porto, Portugal
[3] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC) Porto, Portugal

Abstract

Chest radiography is a widely used imaging exam for pathology diagnosis globally. However, the interpretation of these images may vary between radiologists and is a time-consuming task. This fueled the development of automatic tools for chest radiography interpretation, and within the tasks needed to reach that end, lung field segmentation is a vital one. In this study an automatic method for lung segmentation was developed using a less complex variation of the U-Net architecture. Three datasets were used for this purpose: JSRT, Shenzhen and Montgomery. In addition, two different loss functions were studied: Dice Loss (DL) and Tversky Focal Loss (TFL). The DL model obtained a Dice Score Coefficient (DSC) of $95.34 \pm 3.25\%$ while the TFL model performed at a DSC of $94.74 \pm 2.85\%$, both models being mainly able to correctly predict the lung pixels in either pathological and non-pathological images, but having the worst predictions when a pathology is present. With a higher DSC, the DL model proved to be a better option. In conclusion, this study proves yet again the importance of the U-Net architecture in image segmentation, especially in biomedical images.

1 Introduction

Chest radiography (CXR) is one of the most common imaging examinations globally, playing an essential role in screening, diagnosis, and disease management.
The interpretation of CXR images by radiologists is a time-consuming task, with significant variability between radiologists.This has fueled the development of automatic tools for CXR classification, pathology detection and the segmentation of anatomical structures.
Within the different tasks needed for automatic CXR interpretation, lung field segmentation is a vital task to identify the region of interest for the detection of multiple pathologies and the suppression of false positives, as well as for computing the cardiothoracic ratio, an essential step for cardiomegaly detection. Normally, deep learning methods that rely on convolutional neural networks (CNN's) are used for this purpose.
As such, the aim of this study was to develop an automatic method for lung field segmentation, based on a U-Net [5] architecture, and validate it on multiple datasets, combining both normal and pathological cases.

2 Methods

2.1 Datasets

Three public datasets of CXR images were used: JSRT [6], Shenzhen [2, 3] and Montgomery [3]. The total number of images (with their respective masks) and their distribution regarding the presence or absence of a pathology is shown in the following table:

Table 1: Distribution of images per dataset and by presence or absence of pathology.

Datasets	Images	Pathologic	Non Pathologic
JSRT	247	154	93
Shenzhen	561	282	279
Montgomery	138	58	80
Total	946	494	452

Even though the datasets represent the same type of image (CXR), it is important to note that there are small differences between them. The pathology present in the Montgomery and Shenzhen dataset is tuberculosis while the JSRT images possess lung nodules that can be either malignant or benign. Also, regarding the Shenzhen dataset, the lung masks available were performed by Engineering students and teachers, rather than radiologists, which compromises their accuracy.

2.2 Model Architecture

The architecture used is a slight variation of the U-Net [5], where the difference from the original architecture is the number of channels, which were halved. With this change, the number of trainable parameters was reduced from 30 million to 7 million. This change reduces the complexity of the model which diminishes the time needed for training while maintaining good results. Also, a lower complexity is better since the amount of data available for training is small.

2.3 Evaluation Metrics

In order to evaluate the models, three metrics were used: Dice Similarity Coefficient (DSC), Sensitivity (Sns) and Specificity (Spc).

$$DSC = \frac{2TP}{2TP+FP+FN},\ Sns = \frac{TP}{TP+FN},\ Spc = \frac{TN}{TN+FP} \quad (1)$$

The DSC measures the overlap between the ground truth and predicted mask. Sensitivity is the percentage of lung pixels that were correctly classified as such. Finally, specificity is the percentage of non-lung pixels that were correctly classified as such.

2.4 Loss Functions

The effectiveness of two different loss functions was studied: Dice Loss (DL)[7] and Tversky Focal Loss (TFL) [1].

$$DL = 1 - \frac{2TP}{2TP+FP+FN} \quad (2)$$

$$TFL = (1-TL)^{\gamma},\ \ where\ \ TL = \frac{TP}{TP+\alpha FN+\beta FP} \quad (3)$$

The Tversky Loss (TL) uses two parameters, α and β, where $\alpha+\beta = 1$. By setting the value of $\alpha > 0.5$, the false negatives are more penalised, which is useful when the dataset has a class imbalance and therefore can provide better results with an additional level of control when compared to the DL.

The TFL is a generalisation of the TL. The parameter γ controls the non-linearity of the loss. When class imbalance exists, the TFL becomes useful as the value of γ becomes higher, since it forces the model to focus on misclassified examples, with lower loss being propagated from correctly classified examples. According to the study mentioned in [4], the optimal value for γ is 2.

2.5 Training

Different values were tested for the learning rate using the Adam optimiser. The optimal value for both DL and TFL is 1×10^{-5}. With higher values (1×10^{-3} and 1×10^{-4}) it is not possible to train the model because of overshooting the minimum of the loss function which causes divergent behaviors. With a lower learning rate such as 1×10^{-6} the training is slower with no noticeable benefits.

Two models were trained for 150 epochs, with the aforementioned loss functions, The parameters chosen for the TFL were: $\alpha = 0.7$ (therefore, $\beta = 0.3$) and $\gamma = 2$. The α value was chosen due to the class imbalance regarding lung and non-lung pixels.

The models were trained using 5-fold cross validation. The distribution of the images were made the following way: 60% for the train set, 20% for the validation set and the final 20% for the test set. All images were resized to 512×512 pixels and normalized to an equal intensity range.

3 Results and Discussion

The DL provides a model with a better DSC and specificity, while the TFL generates a model with higher sensitivity. This difference is caused by the α parameter, that takes into account the class imbalance present in every CXR image, where the number of non-lung pixels is higher than the number of lung pixels. Therefore this model has a higher ability to correctly predict true positives, but in general has a worse performance with a lower DSC.

Table 2: Dice Scores, Sensitivity and Specificity results for the DL and TFL models

	DSC(%)	Sensitivity(%)	Specificity(%)
Dice Loss	**95.34± 3.25**	95.16± 5.81	**98.57 ± 1.09**
Tversky Focal Loss	94.74± 2.85	**96.98± 4.57**	97.37± 1.42

Regarding the presence of a pathology in the CXR, the TFL model shows no performance difference between predictions for non-pathological and pathological CXR, having almost the same dice score for both cases. The DL model is clearly better at predicting non-pathological CXR. This evidence is further proof of the effect that the γ parameter has in shifting the focus of the model into learning how to correctly predict harder examples. Nevertheless, the DL model achieved better results for both cases.

Table 3: Dice Scores obtained regarding CXR with and without pathology.

	Non Pathologic(%)	Pathologic(%)
Dice Loss	95.54± 3.03	95.14± 3.43
Tversky Focal Loss	94.76± 3.01	94.72± 2.70

The results per dataset show that both models have better performances predicting images from the Montgomery dataset, while having lower scores regarding the JSRT and Shenzhen datasets. The lower results in the Shenzhen dataset are probably caused by the low quality of the lung masks, since they were not performed by radiologists, as mentioned previously in section 2.1.

Table 4: Dice Scores obtained for each dataset.

	JSRT(%)	Shenzhen(%)	Montgomery(%)
Dice Loss	94.40±4.06	95.37±2.87	96.86±2.35
Tversky Focal Loss	94.87±2.89	94.29±2.84	96.32±2.09

In Figure 1, the best and worst predictions obtained are depicted. In the best scenario, the prediction almost exactly matches the ground truth. However, in the worst case, most of the lung is not noticeable probably due to a case of tuberculosis. Here, the model struggles to correctly predict the full extent of the lungs. As a general rule, the worst predictions happen in cases where a pathology is present.

Also to note at the bottom left corner of the image, a small group of pixels wrongly predicted as lung pixels. This situation is common in the worst predictions. A possible solution would be to post-process the predictions using, for example, connected-component analysis. This way, all the small groups of wrong predictions are removed and a better dice score is obtained.

4 Conclusion

This work proves yet again the importance of the U-net architecture in image segmentation, allowing the development of two automatic lung field segmentation models in CXR images. This study is another step towards the development of automatic tools to interpret CXR, helping to highlight regions of interest for the detection of multiple pathologies. However it is important to note that this research was done with a limited amount of time and computing power. Most of the decisions regarding batch size and learning rate were done in order to minimize the model training time while maintaining performance.

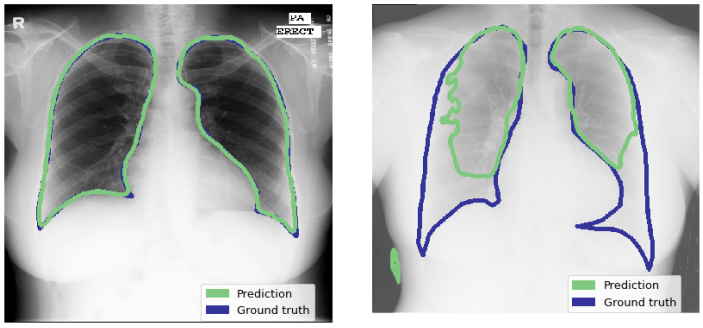


Figure 1: Best (left) and Worst (right) predictions. The Blue contour represents the Ground Truth and the Green contour represents the prediction. Both predictions are from the DL model.

Keeping the same architecture, data and loss functions, better results could possibly be obtained by using a higher batch size, higher number of epochs and a lower learning rate. In addition, trying different architectures, using a pre-trained U-net or different loss functions would be other options to further improve this work and obtain better results. Finally, the problematic of having a small number of images to train could be solved via data augmentation (such as contrast and zoom).

Acknowledgments

This work was funded by the ERDF - European Regional Development Fund, through the Programa Operacional Regional do Norte (NORTE 2020) and by National Funds through the FCT - Portuguese Foundation for Science and Technology, I.P. within the scope of the CMU Portugal Program (NORTE-01-0247-FEDER-045905) and UIDB/50014/2020. J. Rocha and S. Pereira are supported by FCT grant contracts 2020.06595.BD and 2020.10169.BD respectively.

References

[1] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 683–687. IEEE, 2019.

[2] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590, 2013.

[3] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[6] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[7] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.

A semi-supervised approach for colorectal cancer diagnosis from H&E whole slide images

Sara P. Oliveira[1,2] (sara.i.oliveira@inesctec.pt)
Pedro C. Neto[1,2] (pedro.d.carneiro@inesctec.pt)
Diana Montezuma[3] (diana.felizardo@impdiagnostics.com)
Liliana Ribeiro[3] (liliana.ribeiro@impdiagnostics.com)
Ana Monteiro[3] (ana.monteiro@impdiagnostics.com)
Isabel M. Pinto[3] (isabel.macedo.pinto@impdiagnostics.com)
Jaime S. Cardoso[2,1] (jaime.cardoso@inesctec.pt)

[1] INESC TEC
Porto, Portugal
[2] Faculty of Engineering
University of Porto, Porto, Portugal
[3] IMP Diagnostics
Porto, Portugal

Abstract

Pathology labs are evolving to digital workflows, with tissue samples being digitised and evaluated on screen, allowing the development of advanced image processing techniques based on artificial intelligence (AI). Nevertheless, despite colorectal cancer (CRC) being the second deadliest cancer type worldwide, the application of AI for CRC diagnosis, particularly on whole-slide images (WSI), is still a young field. In this work, we propose a semi-supervised approach, based on multiple instance learning, developed on the CRC dataset (1133 H&E WSI), using 100 annotated and 774 non-annotated slides for training, and 259 slides for evaluation. The proposed method attained 88.42% classification accuracy, a Quadratic Weighted Kappa of 0.863 and 95.74% sensitivity.

1 Introduction

Over the last decade, the advent of digitised tissue samples to Whole Slide Images (WSI), the wider adoption of digital workflows in pathology labs, and the consequent availability of more data, combined with a shortage of pathologists, enabled the evolution of the computational pathology field with the integration of automatic image analysis into clinical practice, mainly based in Artificial Intelligence (AI) methodologies [1, 6]. Researchers have been exploring the implementation of computer-aided diagnosis (CAD) systems for several different tasks, such as detection, grading and/or segmentation of lesions.

While colorectal cancer (CRC) can be detected by imaging techniques, further diagnosis is always based on samples obtained from biopsies and assessed by pathologists. Regarding the development stage, these samples can be stratified from non-neoplastic to invasive carcinomas, from the lowest to the highest level of cancer progression, respectively. Although this grading is somewhat subjective [7], the most recent guidelines recommend surveillance for polyps with high-grade dysplasia regardless of their size [4, 5]. Thus, this remains a very important task for pathologists when assessing colorectal tissue samples.

Despite the ever-growing number of publications of machine learning (ML) methods applied to CAD systems, there is a dearth of published work for the task of joint detection and classification of colorectal lesions from WSI, lagging colorectal cancer (CRC) behind pathologies such as breast and prostate cancer. Furthermore, a significant amount of the work developed does not use the entire WSI but instead uses crops and regions of interest extracted from these images. While these latter works show significant results, the applicability in clinical practice is limited.

In this paper, we propose a method that combines weakly and supervised learning methods to diagnose CRC from Haematoxylin-Eosin (H&E) stained slides, without the need of an extended annotated dataset.

2 Methodology

Due to the high dimensionality of WSI (usually over $50,000 \times 50,000$ pixels), each slide is firstly decomposed into small tiles (512×512 pixels). This tissue sampling is based on an Otsu's thresholding mask, that clearly separates the tissue from the background. Tiles that are completely within the foreground are then fed to the classifier (Figure 1).

Traditional supervised learning techniques would require labelled tiles but cancer grading aims to classify the entire WSI and not the individual tiles. Also, tile labelling would easily become a significant workload for pathologists. Therefore, techniques such as multiple instance learning (MIL), can be adapted to computational pathology problems, since it only requires slide level labels, converting the original supervised problem into a weakly-supervised one.

The nature of CRC grading allows the implementation of MIL knowing that, if a slide is classified as class Y, none of its tiles belong to a more severe class (higher grade) and at least one tile belongs to class Y. Therefore, using this concept, we propose a workflow (Figure 1) based on the work of Campanella *et al.* [2] with several adaptations:

(**a**) **Ordinal classes:** in order to bring clinical background to the premises of MIL, slides' classes must not be seen as independent and their relation must be modelled. For instance, normal tissue is closer to low-grade lesions than to high-grade dysplasia, which implies an order in the classes;

(**b**) **Removal of recurrent aggregation:** The original approach includes a Recurrent Neural Network (RNN) to aggregate individual tiles into a final prediction, but in the CRC problem, all tests showed degraded performance. Thus, this step was removed and the slide prediction was based on one tile;

(**c**) **Tile ranking using the expected value:** To select the most representative of potentially thousands of tiles per slide, the backbone network is firstly used to compute the outputs of each tile and the expected value is then computed from these outputs;

(**d**) **Loss function:** Due to the ordinality of classes, the minimisation of the cross-entropy fails to fully capture the model's behaviour. Thus, in an attempt to model the unequal distances between classes, the model is optimised to minimise a loss function based on the Quadratic Weighted Kappa (QWK) [3].

2.1 Dataset

The CRC dataset [8] contains 1,133 colorectal biopsy and polypectomy slides, scanned at 40×, classified into three categories: non-neoplastic

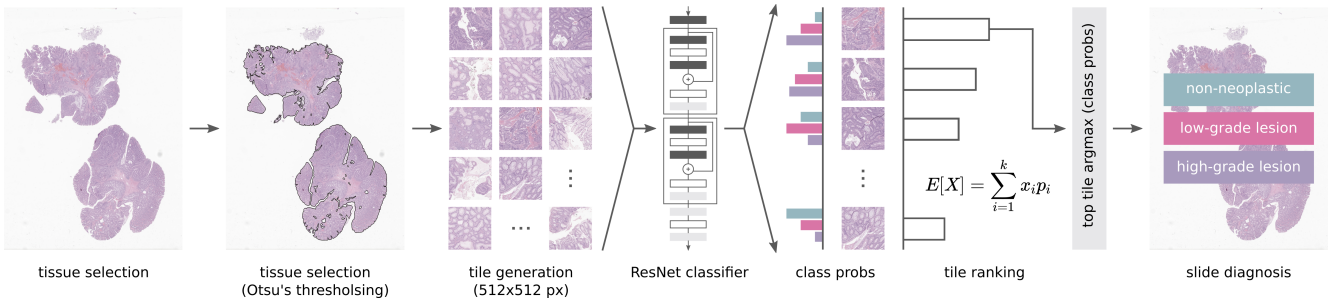


Figure 1: Proposed workflow for colorectal cancer diagnosis on H&E stained whole-slide images.

(normal colorectal mucosa, nonspecific inflammation or hyperplasia), low-grade lesions (conventional low-grade adenomas), and high-grade lesions (conventional adenomas with high-grade dysplasia and invasive adenocarcinomas). All cases were reviewed and labelled by two pathologists and reevaluated by a third pathologist, in case of diagnosis disagreement. From the dataset, a set of 100 samples were manually annotated with region marks, including the abovementioned categories.

2.2 Training details

The proposed model includes a ResNet-34 as the backbone and the experiments were conducted with a batch size of 32 tiles of 512×512 pixels that include 100% of tissue. The network was optimised with the Adaptive Moment Estimation (Adam) algorithm, with a learning rate of 1×10^{-4}, and mixed-precision from the Pytorch available package. As for hardware, the experiments were conducted using an Nvidia Tesla V100 (32 GB) GPU.

3 Results & Discussion

Besides testing the weakly supervised model explained above, the experiments conducted in this work also explore the potential of a subset of annotated data in order to improve the performance of the overall MIL method. Table 1 shows the results of training the model only with the annotated cases, with all cases using only the slide label as reference, and with all cases using a pre-training stage with the annotated subset. As can be noted, there are notable performance gains in both the accuracy and the QWK score as the number of training samples increases. However, perhaps the most exciting performance gain is related to the pre-training of the backbone network on the annotated subset for only two epochs before the start of the MIL training. This experiment is able to outperform the best epoch of the experiment without pre-training in only 7 epochs, in other words, 12 hours of training, with 84.94% accuracy and 0.803 QWK score. Moreover, these values kept increasing until the last training epoch, reaching an accuracy and QWK score of 88.42% and 0.863, respectively. The final results presented in Table 1 can be extended with a sensitivity to high-grade lesions of 93.33% and 95.74% for the last two entries respectively. The model was trained with a set of 874 samples (100 annotated and 774 non annotated slides), whereas the test set had 259 WSI.

Table 1: Performance of the model on the different experiments.

Dataset	Pre-train	QWK score	Accuracy
CRC Annotated (n=100)	No	0.583	75.00%
CRC All (n=1,133)	No	0.795	84.17%
CRC All (n=1,133)	Yes	**0.863**	**88.42%**

The results shown in Figures 2 and 3, respectively for the QWK and the accuracy, are representative of the gains that both the number of samples and the use of annotations bring to the model. Moreover, the use of annotations appears not only to speed up convergence at high values, but also to increase the model's ability to learn at further epochs.

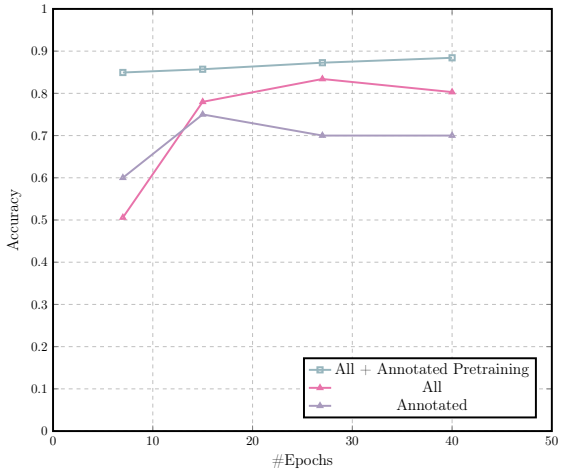


Figure 2: Quadratic Weighted Kappa score evolution (test set).

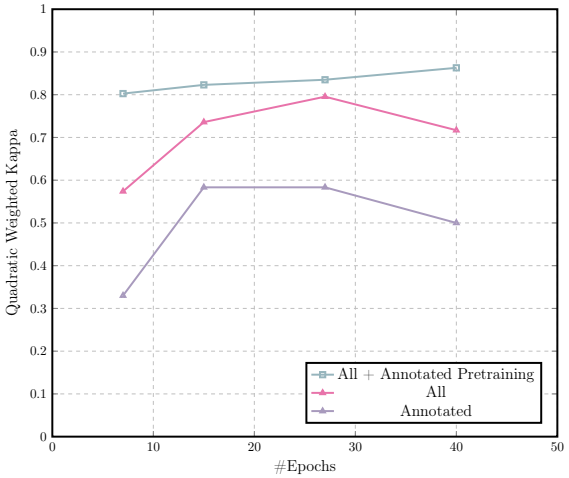


Figure 3: Accuracy evolution (test set).

4 Conclusions

Despite the growing popularity of computational pathology, there are still relatively few published works on CRC diagnosis from entire WSI. Moreover, the reported results are based on relatively small and private datasets. Nonetheless, the construction of larger histopathology datasets with extensive annotations is not an easy and expeditious task. Hence, weakly supervised learning and models that could leverage partially annotated datasets can be explored for such task. In this work we proposed a semi-supervised framework based on MIL, that achieves good results, with only 10% of annotated data, that are consistent with other works for CRC diagnosis, even the most supervised ones, indicating the promising performance of the method. Nevertheless, further efforts can be devoted to performance improvement, mainly regarding the dataset increase and variability, and other approaches to leverage the supervised pre-training stage. Also, interpretability and explainability should be explored in order to inform pathologists about the spatial location that was most responsible for the diagnosis and to explain the reasons behind the prediction.

Acknowledgements This work is financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0247-FEDER-045413 and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within the PhD grant SFRH/BD/139108/2018.

References

[1] B. Acs, M. Rantalainen, and J. Hartman. Artificial intelligence as the next step towards precision pathology. *J Intern Med*, 288(1):62–81, 2020. doi: https://doi.org/10.1111/joim.13030.

[2] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.*, 25(8):1301–1309, 2019. doi: https://doi.org/10.1038/s41591-019-0508-1.

[3] Jacob Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. doi: https://doi.org/10.1037/h0026256.

[4] Samir Gupta, David Lieberman, Joseph C. Anderson, Carol A. Burke, Jason A. Dominitz, Tonya Kaltenbach, Douglas J. Robertson, Aasma Shaukat, Sapna Syngal, and Douglas K. Rex. Recommendations for follow-up after colonoscopy and polypectomy: A consensus update by the us multi-society task force on colorectal cancer. *Gastrointestinal Endoscopy*, 2020. doi: https://doi.org/10.1016/j.gie.2020.01.014.

[5] Cesare Hassan, Giulio Antonelli, Jean-Marc Dumonceau, Jaroslaw Regula, Michael Bretthauer, Stanislas Chaussade, Evelien Dekker, Monika Ferlitsch, Antonio Gimeno-Garcia, Rodrigo Jover, Mette Kalager, Maria Pellisé, Christian Pox, Luigi Ricciardiello, Matthew Rutter, Lise Mørkved Helsingen, Arne Bleijenberg, Carlo Senore, Jeanin E van Hooft, Mario Dinis-Ribeiro, and Enrique Quintero. Post-polypectomy colonoscopy surveillance: European society of gastrointestinal endoscopy guideline - update 2020. *Endoscopy*, 52 (8):687–700, 2020. doi: https://doi.org/10.1055/a-1185-3109.

[6] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016. doi: https://doi.org/10.1016/j.media.2016.06.037.

[7] Dipti Mahajan, Erinn Downs-Kelly, Xiuli Liu, Rish K. Pai, Deepa T. Patil, Lisa Rybicki, Ana E. Bennett, Thomas Plesec, Oscar Cummings, Douglas Rex, and John R. Goldblum. Reproducibility of the villous component and high-grade dysplasia in colorectal adenomas <1 cm: Implications for endoscopic surveillance. *American Journal of Surgical Pathology*, 37(3):427–433, March 2013. doi: https://doi.org/10.1097/PAS.0b013e31826cf50f.

[8] Sara P Oliveira, Pedro C Neto, João Fraga, Diana Montezuma, Ana Monteiro, João Monteiro, Liliana Ribeiro, Sofia Gonçalves, Isabel M Pinto, and Jaime S Cardoso. Cad systems for colorectal cancer from wsi are still not ready for clinical acceptance. *Scientific Reports*, 11(1):1–15, 2021.

Pest detection: Can we beat the technicians?

Bruno Cardoso[1]
badscc@student.dei.uc.pt
Abdellahi Brahim[1]
abdellahi@student.dei.uc.pt
Catarina Silva[1]
catarina@dei.uc.pt
Joana Costa[12]
joanamc@dei.uc.pt,joana.costa@ipleria.pt
Bernardete Ribeiro[1]
bribeiro@dei.uc.pt

[1] CISUC - Department of Informatics Engineering
University of Coimbra, Portugal
[2] School of Technology and Management
Polytechnic Institute of Leiria, Portugal

Abstract

We are dependent on plants to eat, produce medicines or cosmetics. Pests affect the quantity and quality of these cultivated plants by restricting their growth and affect the agriculture revenue by spreading through crops. So, pests monitoring, under the integrated pest control, is essential to maximize the quality and quantity of agriculture production. This monitoring is mostly done by counting the pests in traps, relying on the correct detection of each insect by a technician. In this work, we propose replace part of the technicians work by object detection models for detecting and counting whiteflies, in yellow sticky traps. Results suggest that object detection models can improve whitefly monitoring.

1 Introduction

In the 90s, a whitefly outbreak in the United States made 500 million of dollars in losses. In South America, in regions that already adopted whitefliy control systems, the effects were not felt as much and there was an annual benefit of 2400 dollars per hectare [1]. Today the effect of whiteflies continues to be felt with producers having to adopt the integrated pest control (IPC) to minimize the losses. The IPC starts by preventing the pests, for example, keeping the crop clean. Then, monitoring the pests, counting the number of pests trapped and setting a threshold for this counting. Lastly, if the count exceeds the threshold, chemical control can be used through pesticides, or biological control, through predatory insects. So, a correct monitoring can reduce the use of pesticides while improves the quality of the crops and maximizes the crop revenue.

Technicians carry out IPC monitoring by going to the crops to detect and count the pests present in the traps. This type of monitoring has low accuracy because, for each trap, the technicians only count them in a defined area and extrapolate that count to the rest of the trap. It is expensive because it requires many hours of skilled labour. It also requires extensive and time-consuming work because each trap can contain hundreds of similar insects of different species, making it difficult to classify the insect as a pest or not a pest.

We propose improve the IPC monitoring using object detection models to replace part of the technicians work and prevent them from going, so often, to crops to detect and count the pests. To archive that we use annotated yellow sticky traps images to train object detection models to detect whiteflies. After training, those models can be used, in almost any device, to make whitefly detection and its counting less costly, faster and more precise.

The rest of the paper is structured as follows. Section 2 introduces background on integrated pest control, Section 3 formulates the object detection problem. Section 4 describes the proposed approach and, Section 5 details the experimental setup. Finally, Section 6 concludes and delineates future research lines.

2 Background

The whitefly (WF) is a type of pest that affects, among others, tomato crops. There are several species of whiteflies, visually similar, such as *Trialeurodes vaporariorum* and *Bemisia tabaci*. The evolution of the whitefly starts in the egg, then passes on to nymph and ends up as an adult. Detection at an early stage makes monitoring more efficient.

Yellow sticky traps are commonly used for whitefly monitoring. Usually, this trap consists of a rectangular piece of plastic with a sticky adhesive on the surface to trap small insects. In addition to the yellow colour, used to attract whiteflies, it can also have pheromones. Figure 1 shows part of a Yellow sticky trap dataset image, with nymphs and adult whiteflies.

Figure 1: Yellow sticky trap dataset image

3 Object detection problem

The object detection problem can be decomposed as a classification and regression problem. The classification problem aims to classify the detected object as a whitefly or background of the trap. The regression problem aims to predict the coordinates of the detected object.

The implementation can be defined as one-stage models, like YOLO, or two-stage models, like Region Based Convolutional Neural Networks (R-CNN). The training of two-stage models can be decomposed into 6 steps (as shown in Figure 2):

1. **Set *anchor boxes***: Cover each pixel of a *feature map* by different *anchor boxes* (rectangles of different sizes, similar to annotations, anchored to the center of each pixel).
2. ***Anchor box* classification**: Extract the pixels inside each *anchor box* and classify them as an object (WF) or not (background).
3. **Object classification**: If the classification return an object, classify the object as WF or background.
4. **Coordinates regression**: If the classification returns a WF, predict the coordinate difference (dx,dy, dh, dw) between the *anchor box* and the annotation (it helps in finding the best fitted anchor boxes).
5. ***Anchor boxes* removal**: Remove the overlapping *anchor boxes* that worst fit the same annotation and keep the best ones (Best WF).
6. ***Information loss***: Compare the remaining *anchor boxes* with the annotations, obtain the *information loss* (in classification and regression) and restart the training steps while the *anchor boxes* are not similar to the annotations.

The training of one stage models doesn't have the **Object Classification** step. They immediately classify the information inside each *anchor box* as being whitefly or background. Once trained, the models can be used to detect whiteflies in new images.

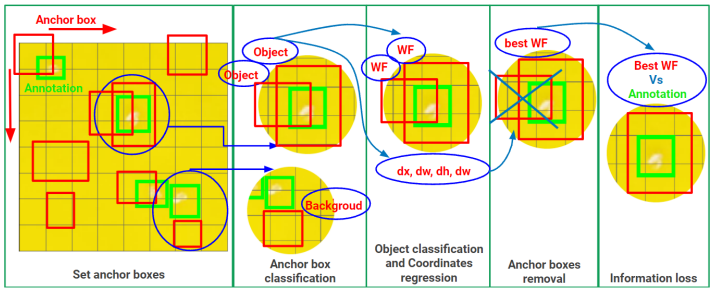


Figure 2: Object detection steps

4 Proposed approach

The main focus of this work is to use object detection models to detect and count whiteflies in trap images. The proposed approach passed through:

1. **Dataset selection**: Choose a public dataset and perform dataset reduction and augmentation.
2. **Model selection**: Choose four object detection models with different architectures and sizes, define the evaluation metrics, compare the performance between those models and choose the best one.
3. **Model validation**: Validate the detections of the chosen model with a technician.
4. **Final results**: Use the validated detections to complete the dataset annotations and evaluate the final model.

This approach allows the evaluation of different object detection models implementations, the trade-off between different evaluation metrics, overcome the lack of annotations and obtain a validated model.

5 Experimental setup

5.1 Dataset

The images used belong to a dataset [2] that contains 284 images with two types of whiteflies (both with WF annotation) and the predatory insects Macrolophus (MR) and Nesidiocoris (NC). This dataset contains 5611 annotated instances of Whiteflies, 1341 annotated instances of Macrolophus, 511 annotated instances of Nesidiocori, and some unannotated WF instances. From this dataset we created the fallowing datasets:

- **"Whitefly dataset":** We removed from [2] the MR and NC annotations, images that contained annotations that did not overlap with the whiteflies, and images that did not contain whiteflies. This dataset contains 174 images with 4940 WF annotations and some unannotated WF. We divided this dataset into a training subset with 139 images and a test subset with 35 images.
- **"Whitefly augmentation datasets":** We created eight datasets adding, for each one, 138 artificial images to the training subset of the "Whitefly dataset" and maintaing the 35 images of the test subset. Each one of the eight datasets contains one type of artificial images (generated by Generative Adversarial Networks (GANs) or geometric transformations like rotation, inversion, or translation).
- **"Whitefly reduction dataset":** We removed, from the "Whitefly dataset", 27 images with unannotated WF instances, but we did not remove images that could change the dataset variability (we kept images with different background colours, whitefly stages and other insects). This dataset contains 147 images, with 117 images belonging to the training subset and 30 images to the test subset.

5.2 Models and Evaluation Metrics

To validate our approach, we chose four object detection models to test on the datasets described above: YOLOv5 (XLarge) and YOLOv5 (Small), which were made public in June 2020. Scaled-YOLOv4 (Large) with an article published in November 2020 and best results in the COCO dataset. And finally, the R-CNN with an article published in 2017, and sometimes with the best results in detecting small objects [3,4].

YOLO models allow testing different depths of the same architecture, from Tiny to XLarge, changing only the number of layers and neurons. This allows testing the state of the art of object detection and evaluating the trade-off between precision and memory consumption or detection time. The R-CNN allows the testing of a two-stage model and compares performance with one stage models.

To evaluate those object detection models, we considered 4 metrics: the *mean average precision* (mAP) being one of the most used metrics to assess the precision of object detection models. As mAP is affected by the lack of annotations in the dataset, we also considered the metric number of *Detected WF*, whether annotated or not, which allows the assessment of which model detects more whiteflies, while considering detected unannotated whiteflies as false positives (FP). We have also considered the metrics, *Memory* consumption and average detection *Time* for an image (using the NVIDIA Tesla T4 GPU).

6 Experimental results

We started by testing the models in the "Whitefly dataset" and got the best result with the YOLOv5 (XLarge) model, with a mAP of 80.00%. To improve the results, we tested the YOLOv5 (XLarge) model in the "Whitefly augmentation datasets", but as explained in [2], using geometric transformations to increase the number of images also increases the number of unannotated WF and using GANs decreases the quality of the images.

Therefore, to remove the effect of unannotated whiteflies, we have tested the models on the "Whitefly reduction dataset". Figure 3 shows that the YOLOv5 (XLarge) model obtained the best results detecting 1280 WF, with 85.10% of mAP, an average detection time of 0.13 seconds and consuming 244 MB of memory.

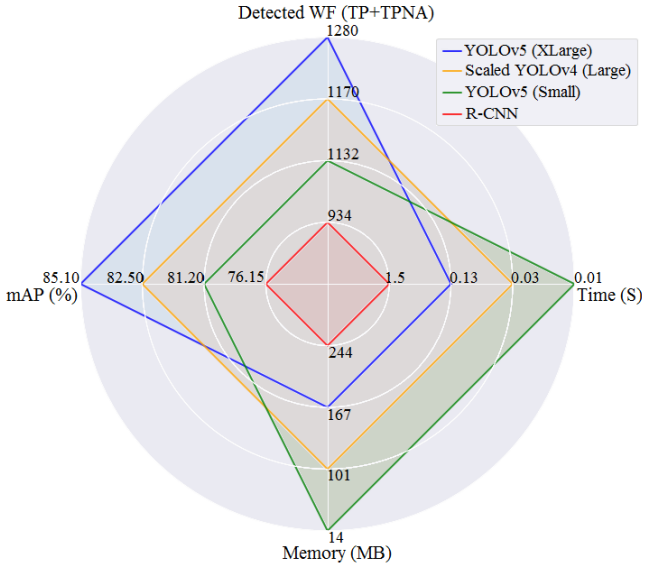


Figure 3: Models performance

Thereafter, we used the YOLOv5 (XLarge) to detect whiteflies in images with unannotated whiteflies, and we submitted those detections to be validated by a technician. From the Table 1, it can be seen that the YOLOv5 (XLarge) model detected 67 unannotated and validated whiteflies (TPNA - True Positive Not Annotated) in the test subset of "Whitefly reduction dataset" improving the detection by 2.20% in relation to the annotators.

TP	TPNA	TN	FP	FN
1141	67	166	21	41

Table 1: YOLOv5 (XLarge) in "Whitefly reduction dataset"

Finally, we used the validated detections to complete the dataset annotations. The total annotations passed from 4939, as described in section 5.1, to 5863 (more 18.71%) and, the mAP improved to 89.7%.

7 Conclusions and future work

We presented a method that uses object detection models for detecting and counting whiteflies. This method consisted of testing four object detection models in a dataset that contained many unannotated whiteflies. The R-CNN model had the worst performance, YOLOv5 (small) was the fastest and Scaled-YOLO (large) had the best trade-off between mAP and speed. The YOLOv5 (XLarge) model had the highest mAP and detected more whiteflies than the dataset annotators. After validating the YOLOv5 (XLarge) detection's and using them to complete the dataset annotations, we got a mAP of 89.70%.

Future work includes improving the creation of artificial image datasets using GANs and implement a remote monitoring system. GANs allows the creation of datasets without unannotated whiteflies or datasets with other types of pests and backgrounds. The remote monitoring system will be made, programming small cameras to collect daily images from the traps and to send those images to the cloud, where the object detection models will count the detected pests, and send a report to the technicians.

References

[1] Oscar Ortiz. Tropical Whitefly Project Progress Report. Impact Evaluation, 2007.

[2] Ard Nieuwenhuizen, Jochen Hemming, and Hyun Suh. Detection and classification of insects on stick-traps in a tomato crop using Faster R-CNN. The Netherlands Conference on Computer Vision, 2018.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick. Mask R-CNN. ICCV, 2017.

[4] Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao. Scaled-YOLOv4: Scaling Cross Stage Partial Network. arXiv:2011.08036, 2020.

Evaluation of different depth camera technologies in transparent and semitransparent scenes

Eva Curto
evacurto@isr.uc.pt
Helder Araujo
helder@isr.uc.pt

Institute of Systems and Robotics
Department of Electrical and Computer Engineering
University of Coimbra

Abstract

In the last decade, various companies have released different versions of RGB-D sensors, improving their performance at various levels (resolution, frame rate, robustness). These devices can measure depth using one of the following optical technologies: Structured-Light, Active Stereoscopy or Time-of-Flight / Lidar. This paper aims to analyze how these different operating modes of the cameras bias the performance in the depth estimation for specific scenarios with transparency and refraction. We propose an experimental setup involving an aquarium and liquids to study transparency and refraction effects in depth estimation. The evaluation is based on repeatability/precision and distribution of the acquired depth. Point-to-plane distances were also used as a precision metric since the scene includes three different possible planes.

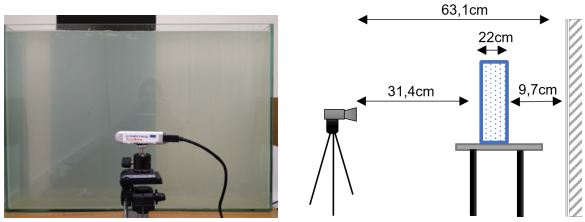
1 Introduction

Consumer RGB-D cameras have been widely used in robotics and computer vision due to their compactness and ability to perform 3D reconstruction in real-time with a frame rate of 15-30 fps and resolutions up to 1280×720 [1], [2], [3]. In this paper, we address three different technologies of range sensing: Structured Light (SL), Active Stereoscopy (AS) and Light Detection and Ranging (LiDAR). Each one was explored using a RealSense camera; more precisely, the models SR305, D415 and L515 were used. SL principle (SR305) uses coded light technology to estimate depth. This means that there is a projector emitting one or more patterns sequentially onto the scene. These patterns are warped by the scene, reflected back and captured by a standard camera. The scene's depth is estimated by comparing the projected patterns and the distorted ones acquired by the camera. Differently, AS (the D415 model) comprises two ordinary cameras and an unstructured light pattern. The depth is estimated by triangulation using the disparity map between the two cameras. The pattern only serves to add artificial features to triangulate it. Finally, the LiDAR L515 which is based on the Time-of-Flight principle (ToF). That is, the system estimates the depth by sending laser pulses on the scene and measuring the time between the laser pulse arriving at the target and return to the receiver. The L515 camera uses an Edge-Emitter Laser for the light source, a MEMS micro-mirror to scan the environment, a photodiode to measure the time of flight and optical lenses to focus the beam. The several differences between these principles lead to distinct performances against possible known error sources such as background light, transparent and reflective materials, multi-path effect and others. In our work, we only handle the transparency/refraction problem. Thus, inspired by the test scenarios of [4] and [5] using semitransparent liquids, we designed our setup also using semitransparent liquids but in an aquarium big enough to fill the whole scene.

2 Methodology

2.1 Experimental Setup

As aforementioned, the core of this work lies in comparing the performance of three different working principles in depth estimation for transparent and reflective targets. For this purpose, Intel RealSense cameras were chosen, specifically the models SR305 (SL) [6], D415 (AS) [7] and the newest L515 (LiDAR) [8], all operating with the Intel RealSense SDK 2. The RealseSense library librealsense2 served as a basis for all the depth acquisition software. For each different acquisition, 100 samples were obtained and saved. In addition to depth, data from the RGB and infrared cameras were also recorded. To ensure a fair comparison, we set the cameras with the same depth resolution 640×480, since it is the unique resolution of the SR305 camera. The remaining streaming settings were used in the default mode except for the L515 camera, where the "Short Range" preset was selected. The acquisitions were made in a darkroom containing an adjustable LED board. In this way, all the acquisitions were performed under constant illumination. Our setup for the transparency experiments was the following: a glass aquarium with dimensions 0.84×0.22×0.58cm (and about 6mm of thickness) on a table in front of a wall and a camera mounted on a tripod pointed to the aquarium. Figure 1(a) shows the back view of the setup and Figure 1(b) illustrates the distances between the camera, the aquarium and the wall.



(a) (b)

Figure 1: The experimental setup for depth acquisition with transparency and refraction: (a) Picture of the back view of the setup with D415 camera; (b) Distances between the camera, the aquarium and the wall. The aquarium is 9.7cm from the wall and the camera is 31.4cm away from the aquarium. So, since the width of the aquarium is 22cm, the distance between the camera and the wall is 63.1cm.

The transparency experiment included several tests explained below:

- *plane:* This is a zero test where we acquire depth images directly from the wall, that is, without any transparency between the camera and the wall. These depth measurements will serve as a reference to the following tests since we don't have ground truth for the depth.
- *empty:* In this test, the aquarium is inserted between the camera and the wall. Therefore, we aim to analyze the influence of the two transparent glass walls of the aquarium (with air in-between) in the depth estimation.
- *water full:* This test introduces another challenging scenario regarding transparency. The aquarium is filled with water (about 95L). Then, between the camera and the wall, we have a first glass wall, water, a second glass wall and air.
- *water milk1/2/3/4:* Set of tests, where the water is dyed with milk to experience different levels of transparency/opacity. *water milk1* is the less opaque of the four, with a milk concentration of 0.03% $(V/V\%)$. Then, *water milk2* with 0.13% $(V/V\%)$ and *water milk3* with 0.24% $(V/V\%)$ Finally, the most opaque solution, *water milk4* with 1.13% $(V/V\%)$. In Figure 1(a), we can see a picture taken during the test *water milk3*.

2.2 Evaluation

In the evaluation, all the 100 acquisitions from each test were used. Having the depth maps, we estimated the average depth for each pixel from the 100 samples. To exclude unreliable estimates, we consider a pixel invalid if it had more than 20 non-positive depth values out of 100 measurements. This data allows us to evaluate the repeatability/precision of the camera and have a general picture of the depth distribution for each camera in every test. The precision was also measured in terms of point-to-plane distance. This was possible since we fitted our data to planes. For

the point-to-plane error evaluation, a rectangular area was segmented in each depth frame (the same area for all 100 samples) regarding the aquarium's central part, avoiding lateral distortions. The average depth for each pixel was also estimated, excluding invalid pixels (as in the previous evaluation). Then, we found existing planes in the data using the M-estimator Sample Consensus (MSAC) algorithm and the point-to-plane distance is then estimated.

3 Results

In this section we show some of the results obtained from the experimental evaluation. In Figure 2 the histograms that show the distribution of the average depth for each one of the seven transparency tests are shown. Each histogram plots the number of points (vertical axis) obtained for each distance (horizontal axis). For each case, we can compare the distributions for the three cameras. Although we did not calculate the average number of invalid pixels, they were added to the histograms in Figure 2 for visualization purposes, allowing us to compare the cameras in what concerns the number of invalid pixels.

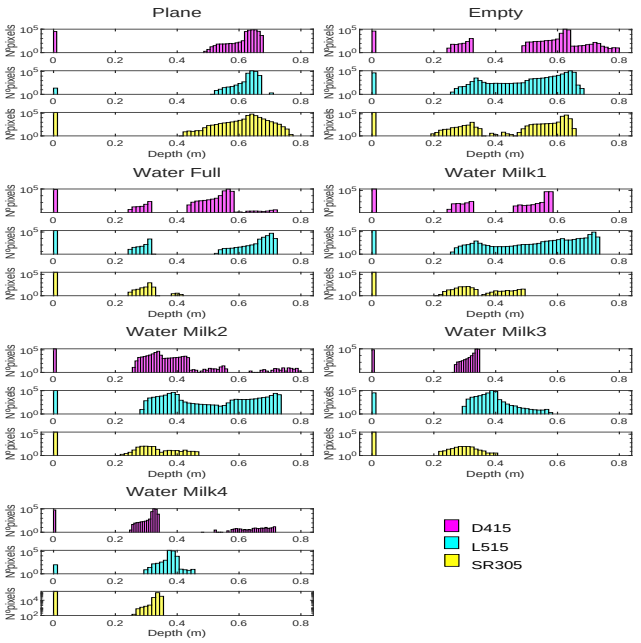


Figure 2: Histograms for each acquisition (allowing a comparison between cameras): each histogram represents the number of pixels (absolute frequency) for each average depth value estimated by each camera. A logarithmic scale is used for all the vertical axes since the data ranges over several orders. For example, the order of the frequency of the invalid depth pixels for the SR305 is much higher than the remaining data.

Figure 3 presents the average point clouds of the evaluated segmented areas colored according to the point-to-plane distance. In some cases, it was not possible to fit the data to planes, hence the empty spaces in *water full*, *water milk1* and *water milk3* tests. For *water milk2* we could not estimate planes for any of the cameras. This can be explained by the distributions of the depth values. That is, if we observe the histograms in the *water milk2* case, these have a scattered depth distribution, meaning that points belonging to a plane are not detected in a sufficiently high number to allow for a robust estimation of a plane. We can also notice that the planes fitted are different depending on the tests. In the first two, the plane corresponds to the wall, in the *water full* and *water milk1* it corresponds to the second wall of the aquarium and in the last two tests it corresponds to the first wall of the aquarium.

4 Conclusion

This paper describes a comparison between different depth cameras technologies in the case of transparency and refraction. We have presented an experimental framework to evaluate the estimation of depth with a scenario composed by a glass aquarium and semitransparent liquids. Com-

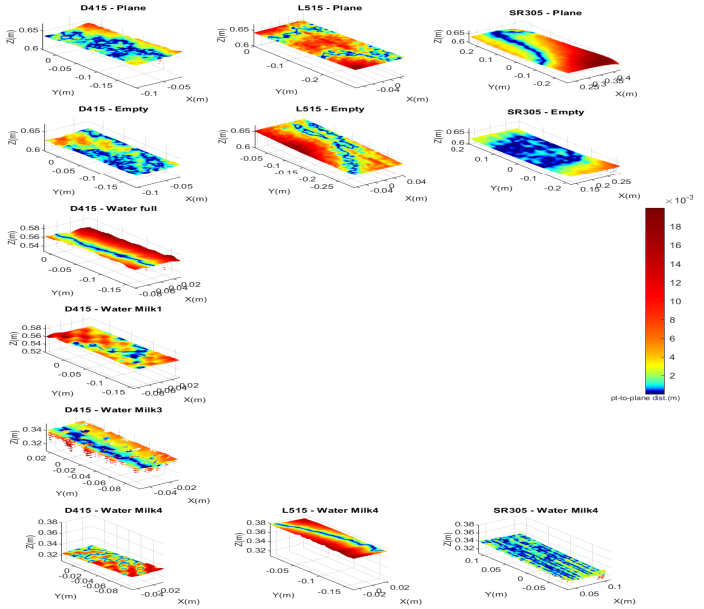


Figure 3: Point clouds of the evaluated segmented areas colored according to the point-to-plane distance.

prehensive results were obtained demonstrating that the D415 camera yields better depth estimates in the case of transparent objects. The current evaluation was limited by the lack of a ground truth depth. This research can be helpful for researchers when considering depth estimation cameras for scenarios with transparency and refraction. These results can be further analyzed taking into account the specific physical principles used to estimate depth.

5 Acknowledgements

This work was partially supported by Project COMMANDIA SOE2/P1/F0638, from the Interreg Sudoe Programme, European Regional Development Fund (ERDF), and by the Portuguese Government FCT, project no. 006906, reference UID/EEA/00048/2013

References

[1] Monica Carfagni, Rocco Furferi, Lapo Governi, Chiara Santarelli, Michaela Servi, Francesca Uccheddu, and Yary Volpe. Metrological and critical characterization of the intel D415 stereo depth camera. *Sensors (Switzerland)*, 19(3), 2019.

[2] Paul L Rosin, Yu-Kun Lai, Ling Shao, and Yonghuai Liu. *RGB-D Image Analysis and Processing*. Springer Nature, 2019.

[3] Michael Zolhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the Art on 3D Reconstruction with RGB-D Cameras. *EUROGRAPHICS 2018*, 37(2):149–154, 2018.

[4] Hamed Sarbolandi, Damien Lefloch, and Andreas Kolb. Kinect range sensing: Structured-light versus Time-of-Flight Kinect. *Computer Vision and Image Understanding*, 139:1–20, 2015.

[5] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Horaud. *Time of Flight Cameras : Principles , Methods , and Applications*. Springer Science & Business Media, 2012.

[6] Intel. Intel ® RealSense ™ Camera SR300 Embedded Coded Light 3D Imaging System with Full High Definition Color Camera. Technical Report June, 2016.

[7] Intel® RealSenseTM. Product Family D400 Series Datasheet. Technical Report November, 2018.

[8] Intel® RealSenseTM. LiDAR Camera L515 user guide. Technical Report July, 2020.

Order is the key: Deep focus assessment in Whole Slide Images

Tomé Albuquerque[12]
tome.m.albuquerque@inesctec.pt
Ana Moreira[1]
anasof3m@gmail.com
Jaime S. Cardoso[12]
jaime.cardoso@inesctec.pt

[1] Faculty of Engineering of the University of Porto
Porto, Portugal
[2] INESC TEC
Porto, Portugal

Abstract

Medical image quality assessment plays an important role not only in the design and manufacturing processes of image acquisition but also in the optimization of decision support systems. This work introduces a new deep ordinal learning approach for focus assessment in whole slide images. From the blurred image to the focused image there is an ordinal progression that contains relevant knowledge for more robust learning of the models. With this new method, it is possible to infer quality without losing ordinal information about focus since instead of using the nominal cross-entropy loss for training, ordinal losses were used. Our proposed model is contrasted against other state-of-the-art methods present in the literature. A first conclusion is a benefit of using data-driven methods instead of knowledge-based methods. Additionally, the proposed model is found to be the top-performer in several metrics. The best performing model scores an accuracy of 94.4% for a 12 classes classification problem in the FocusPath database.

1 Introduction

Digital Pathology images (Whole slide images (WSI)) are about 10X bigger than Radiology images, being over >1 GB in size in most cases, and thus require better storage management through their useful life cycle in clinical workflow. Image quality assessment (IQA) methods are crucial in this cycle; working as a filter in the first stage of acquisition, they can improve storage management. In a second step, they can be used as an optimizer of the decision support systems.

In the medical field image quality assessment methodologies are focused on two distinct processes: on the one hand, the low-level notion of quality which includes the measurement of distortions at a signal level such as blur, noise, compression errors, and other types of distortions; on the other hand, the semantic complex concepts such as the presence/ absence of artifacts (e.g. tissue folds or bubbles, the presence of coloration errors, among others). Thus, the medical image quality assessment is, in most cases, application-specific which requires vast domain knowledge in the respective medical area.

Focus quality assessment (FQA) is fundamental in the normal WSIs acquisition workflow. Hence, it is essential to develop and improve FQA methods capable of improving the quality of acquired WSIs and reduce the acquisition time.

During the acquisition of WSI's by the scanning platforms, focus errors often occur. After the acquisition, manual inspection of the slides is required to infer if it has good quality to proceed with the analysis and diagnostic. The manual inspection of the slides is a time-consuming process and in most cases subjective to individual scores which leads to inter/intra-variability issues between the experts. This way, to automatize this process and improve clinical workflow several focus quality assessment approaches have been developed. They can be divided into Knowledge-based focus quality assessment methods and data-driven focus quality assessment methods.

Knowledge-based FQAs - These methods are based on domain knowledge, with a large presence of this type of algorithms in the literature. Most of this methods require low computational power and are more interpretable; however, compared with the most recent data-driven methods, their performance is relatively low in terms of precision and computation time.

Data-Driven FQAs - In the last few years, there was an advance in data-driven approaches for FQA in WSIs. These type of methods present very good performances in WSIs focus assessment, however, the high computational costs to train these models represent its main drawbacks when compared to Knowledge-based FQAs.

2 Brief Summary

Most of the current approaches to ordinal inference for neural networks are found to not adequately take advantage of the ordinal problem. In the case of FQA for WSIs, all the data-driven approaches present in the literature discard the ordinal information between the different focus classes.

In the present work, a novel model to infer focus quality in WSIs is proposed. This new model is developed on the FocusPath dataset and uses a plethora of deep learning architectures, with several ordinal losses. Our proposal is also compared with the current state-of-the-art approaches and surpasses them in several metrics.

3 Methods

3.1 Data pre-processing

Firstly, the dataset was divided into train, validation, and test subsets (60-20-20%), maintaining the ratio among different classes. To feed the network, it was necessary to resize the patch images from 1024×1024px to 224×224px. To overcome the small amount of data in our dataset, data augmentation was used. Therefore, during the training of the models, a series of random transformations were done in each training epoch for every image.

3.2 Convolutional Neural Networks

A convolutional neural network (CNN) is a class of deep learning algorithms that consecutively apply convolutions of filters to the image. These filters are learned and consist of quadrilateral patches that are convolved across the whole input image – unlike previous fully-connected networks, only local inputs are connected at each layer. Usually, each convolution is intertwined with downsampling operations, such as max-pooling, that progressively reduce the size of the original input image. The final layers are fully-connected and then the final output is processed by a soft-max for the multi-class problems. In this work a model for classification of focus quality in WSIs was trained and tested with seven different convolutional network architectures.

3.3 Losses

In this work six different losses will be evaluated, Cross-Entropy (CE) for the baseline model and five different ordinal losses: Ordinal Encoding (OE), Binomial Unimodal (BU), CO, CO2, and also Ordinal Entropy Loss Function (HO2) [1].

4 Experimental Details

4.1 Dataset

In this work, it was used a public histopathological database named FocusPath dataset, which contains image quality annotations [3]. The FocusPath dataset contains 8640 patches of 1024 x 1024 images. These images were extracted from nine different stained slides from diverse human organs. The original Whole Slide Images were scanned by Huron TissueScope LE1.2. It uses a 40X optics lens at 0.25 μm/pixel resolution. There are 14 absolute z-level scores corresponding to the ground-truth class for the focus level. Due to the low number of examples of images in the more defocused classes (12 and 13), the label was changed to belong

to class 11. This way, the dataset was grouped into 12 different focus classes.

4.2 Training

All the losses were tested on the CNN model represented on Figure 1.

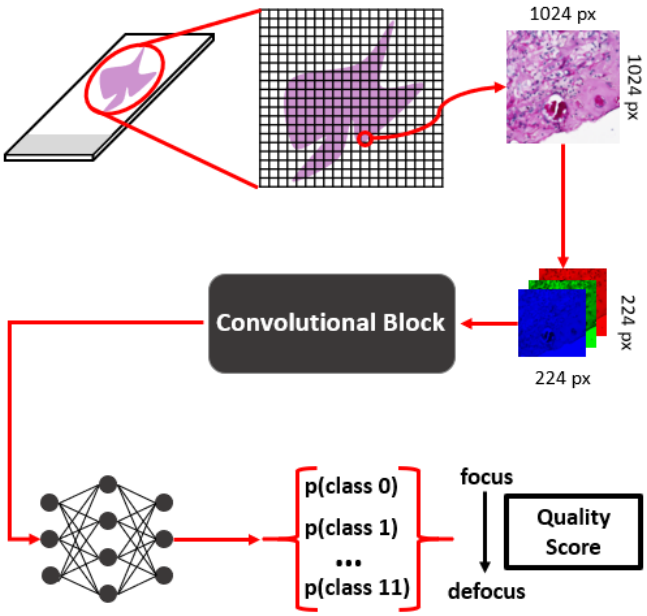


Figure 1: Schematic representation of the model architecture.

The model had as input the RGB WSIs patches from the FocusPath database. The model output consists of the multiclass classification (12 classes) of focus quality (0-focus patch to 11-defocus patch)

During the training initialization, the weights of the architectures previously mentioned were initialized based on ImageNet pre-training. The optimizer used was ADAM, starting with a learning rate of 10^{-4}. The learning rate is reduced by 10% whenever the loss is stagnant for 10 epochs using a specific scheduler. The training process is completed after 200 epochs.

5 Results and Discussion

The performance of the seven different architectures are presented in Table 1, for 12-class classification problem, with the six different learning losses – conventional Cross-Entropy (CE), Binomial Unimodal (BU), Ordinal Encoding (OE), CO, CO2 and Ordinal Entropy Loss Function (HO2). The best models are shown in bold.

Table 1: Results in terms of **Mean Absolute Error (MAE)** for **12 class** problem (lower is better).

	CE	BU	OE	CO	CO2	HO2
alexnet	**0.29**	0.32	0.34	0.37	0.42	0.40
googlenet	0.08	0.13	0.09	**0.08**	0.13	0.22
resnet18	**0.12**	0.13	0.13	0.19	0.19	0.14
mobilenet_v2	0.08	0.18	**0.06**	0.12	0.15	0.16
shufflenet_v2_x1_0	0.16	0.19	**0.16**	0.16	0.25	0.20
squeezenet1_0	0.30	0.29	**0.27**	1.04	0.43	0.42
vgg16	0.14	0.15	**0.11**	0.43	0.17	0.23
Avg	0.17	0.20	0.17	0.34	0.25	0.25
Winners	2	0	**4**	1	0	0

Table 1 presents the results for MAE for 12-class focus classification problem. Regarding this metric, it is possible to infer the difference in the results between nominal and ordinal losses. OE loss achieved the best performance across the different architectures. Across the different architectures, ordinal losses won nominal CE in 5 against 2. This can be explained by the lower role of ordinality in the CE loss. This means that when misclassification occurs, ordinal losses tend to classify the focus quality in the WSI patches as being closer to the real class.

Going a step further and comparing the results with other works featuring focus quality assessment in the FocusPath database, it is clear that the evaluation metrics are above the state-of-the-art results, especially on the Spearman's rank correlation coefficient (SRCC) and Pearson correlation coefficient (PLCC) metrics. Wang et.al [5] present a data-driven work with better results in all the metrics, in contrast to knowledge-based methods (MLV [2] and FQPATH [4]). However, in his approach, the model is fed with random crops (235 x 235 px) of the WSIs patches. Thus, it loses some relevant spatial information about the focus (e.g. when exists in the same patch focus and defocus zones).

This comparison of metrics between our proposed method and the literature works is presented in Table 2.

Table 2: Comparison of our proposal with literature models

	SRCC	PLCC	Time (sec)
MLV [2]	0.8623	0.8528	0.482
FQPATH [4]	0.8395	0.8295	0.269
FOCUSLITENN [5]	0.8931	0.8857	**0.019**
Proposal	**0.9969**	**0.9970**	0.047

6 Conclusions and Future Work

Comparing the different deep learning approaches on WSIs patches, the models trained with ordinal losses achieved better results when comparing with the nominal cross-entropy loss. Thus, a new model has been proposed for multi-class FQA in WSIs based on convolutional neural networks and respecting the ordinal progression among the different focus levels. This new model demonstrated to be competitive with state-of-the-art results and surpass them in some metrics. Furthermore, another important outcome is that those data-driven methods obtained significantly better results than knowledge-based methods in terms of precision and performance.

For future works, since this model is a prototype of an FQA system, the architecture should suffer some changes and multi-processing must be used to reduce the processing time per patch in a real-time acquisition system for WSIs. The implementation of multiple instance learning in training would be interesting to replace the traditional resizing of the images where information loss may occur of some relevant features.

Acknowledgments

This work was partially funded by the Project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership. Tomé Albuquerque was supported by Ph.D. grant 2021.05102.BD, also provided by FCT.

References

[1] Tomé Albuquerque, Ricardo Cruz, and Jaime S. Cardoso. Ordinal losses for classification of cervical cancer risk. *PeerJ Computer Science*, 7:e457, April 2021. ISSN 2376-5992. doi: 10.7717/peerj-cs.457. URL `https://doi.org/10.7717/peerj-cs.457`.

[2] Khosro Bahrami and Alex C. Kot. A fast approach for no-reference image sharpness assessment based on maximum local variation. *IEEE Signal Processing Letters*, 21(6):751–755, 2014. doi: 10.1109/LSP.2014.2314487.

[3] M. S. Hosseini, Y. Zhang, and K. N. Plataniotis. Encoding visual sensitivity by maxpol convolution filters for image sharpness assessment. *IEEE Transactions on Image Processing*, 28(9):4510–4525, Sep. 2019. ISSN 1057-7149. doi: 10.1109/TIP.2019.2906582.

[4] M. S. Hosseini, Jasper A. Z. Brawley-Hayes, Y. Zhang, Lyndon Chan, K. Plataniotis, and S. Damaskinos. Focus quality assessment of high-throughput whole slide imaging in digital pathology. *IEEE Transactions on Medical Imaging*, 39:62–74, 2020.

[5] Zhongling Wang, Mahdi S. Hosseini, Adyn Miles, Konstantinos N. Plataniotis, and Zhou Wang. Focuslitenn: High efficiency focus quality assessment for digital pathology, 2020.

Question Answering from Technical Portuguese Documents

Sara Inácio
sarainacio@student.dei.uc.pt
Hugo Gonçalo Oliveira
hroliv@dei.uc.pt
Catarina Silva
catarina@dei.uc.pt

Universidade de Coimbra
CISUC - Centro de Informática e Sistemas
FCTUC-DEI - Departamento de Engenharia Informática
Coimbra, Portugal

Abstract

Given the relevance of Automatic Question Answering for domain-specific conversational agents, we compare approaches for obtaining answers from a collection of technical documents written in Portuguese. We experiment with traditional IR, fine-tuned neural language models, and the combination of both. Our main conclusion is that a combination of BERT with a IR engine, used for pre-selecting the most relevant documents, retrieves a high percentage of quality answers, whilst not requiring a demanding configuration complexity or great time consumption.

1 Introduction

Dialogue systems have progressed substantially from conversational agents focused on emulating human-to-human conversations to agents capable of answering domain questions, acquired from unstructured sources. Automatic Question Answering (QA) is a task in the scope of Natural Language Processing (NLP) and Information Retrieval (IR) that consists of finding the answers for natural language questions in collections of text. Along with recent advances of these systems, comes the effort of their creation and maintenance, which tends to escalate with the size and diversity of the target domain. In addition, such effort must be applied whenever a new agent is created, or the domain altered, leading to an intensive requirement of both human and financial resources. Available approaches for QA from text range from traditional IR [4], where suitable answers are searched in corpora, to neural approaches [2], where an output is derived from the user's input.

This work explores a set of approaches for QA in Portuguese, in any given domain reflected in an available collection of documents. Those include an IR-based approach, i.e., the text search engine Whoosh, and two types of state-of-the-art transformer-based language models, namely BERT, bidirectional and fine-tuned for extracting answers given a context and a question; and answer-generating GPT2 and GPT3. The goal is to assess the selected approaches, whilst considering relevant aspects such as human labour, time consumption, and answer quality. The upcoming sections describe in detail the explored approaches, as well as the conducted experiments. Lastly, the results and drawn conclusions are presented.

2 Explored Approaches

Explored approaches for Portuguese QA can be divided into three categories: (i) traditional IR; (ii) fine-tuned neural language models; (iii) a combination of (i) and (ii). Following a traditional IR approach, Whoosh[1] is a library that allows to index text and find matching documents based on a given search string. Even though it was not built for QA, if the search terms are in the form of a question, Whoosh will find the document(s) with the most similar text snippet and highlight this snippet, which, according to human intuition, will often work out as an answer. Due to its lower complexity, we see IR as a baseline. Nevertheless, we also explore neural language models that recently became the paradigm for many NLP tasks, including QA. These were GPT2 [6], after downloading the model, and GPT3 [1], through OpenAI's API[2]. In order to generate answers, both were fine-tuned in the collection of documents.

Another well-known transformer-based architecture is BERT [3]. For Portuguese, BERTimbau [7] is a pre-trained BERT model, recently fine-tuned[3] for open-domain Portuguese QA[4]. More precisely, given a textual context and a question, this model returns the span of text that better answers the question. Since the answer could be in any of the documents, one possibility would be to use each document as context, then extract the best answer from each, and finally select the best, e.g. based on an assigned score. However, this would be extremely time-consuming and grow linearly depending on the size of the the collection. Therefore, we propose a combined strategy that makes an initial selection of documents based on a Whoosh index. Only the documents retrieved by Whoosh in the first step are used by BERT as context for answer extraction.

[1] https://whoosh.readthedocs.io/ [2] https://beta.openai.com/ [3] While GPT aims to learn any task unsupervisedly or in a few shots, BERT has to be fine-tuned for the task to perform.
[4] https://huggingface.co/pierreguillou/bert-base-cased-squad-v1.1-portuguese

3 Experimental Setup

Each approach was configured and tested with the same data and setup, i.e., a collection of documents and related questions, with common metrics used to evaluate the answers by each approach.

3.1 Data

We compiled a collection of 25 Portuguese documents about telecommunication equipment. Then, we manually created a set of 67 questions to be answered with information in the collection, and mapped each to the text that would provide its answer.

Since there are similar pieces of equipment, some documents tend to have a similar structure. Thus, to minimise confusion, before indexing or fine-tuning, the filename of each document, often the name of the equipment, was automatically added to every paragraph in the document (see Figure 1). Moreover, for identifying questions and their answers, a `P:` was added before each question and a `R:` before each answer, which follows its question immediately (see Figure 2).

```
Placa ZTE MF65 - O utilitário (ZGPatchForEcmDriverV1.0.6.pkg)
1. Sistema Operativo: OS X Yosemite (10.10).
2. Equipamentos: ZTE MF667, ZTE MF63, ZTE MF65.
3. Sintomas: Depois de instalar o equipamento, o mesmo não é detectado.
4. Solução: Instalar o utilitário fornecido pela ZTE de acordo com os passos descritos neste guia.

Placa ZTE MF65 - Requisitos:
1. Computador ligado à corrente, para garantir um melhor desempenho.
2. Software do equipamento ZTE (router/placa de dados) instalado no computador.
3. Equipamento ZTE (Router/Placa de dados) desligado do computador.

Placa ZTE MF65 - Detalhes de Instalação:
1. Caso o requisito 1 não seja verificado, ligue o carregador do computador à corrente.
```

Figure 1: Example of Document with Identifier

```
P: Quais os requisitos para montar a placa ZTE MF65?
R: Placa ZTE MF65 - Requisitos:
1. Computador ligado à corrente, para garantir um melhor desempenho.
2. Software do equipamento ZTE (router/placa de dados) instalado no computador.
3. Equipamento ZTE (Router/Placa de dados) desligado do computador.
```

Figure 2: Example of Question and Corresponding Answer

As the focus of this work lies in systems capable of adapting to any given Portuguese domain, whether a new one or an adjustment of the current, the domain in the conducted experiments could indeed be any set of textual documents written in Portuguese.

3.2 Configuration

Even though each approach relies on a specific configuration method, ranging from indexation of text to hyperparameter specification or fine-tuning on the domain, all are submitted to the same experiences.

For Whoosh, the collection of documents was first indexed. Each question may then be used as the search string, to retrieve the document(s) with the most similar text span. However, instead of the full document, only this span (i.e., highlighted by Whoosh) was used as the answer. Apart from defining the search string, this IR-based approach supports the definition of different options, such as selecting the token analyzer or specific filters. Upon testing different combinations, we decided to add a 4-gram filter to the analyzer, which breaks individual tokens into groups of four characters when searching for the most relevant document. Even if differences between different configurations were minimal, this configuration was slightly superior in our data.

Despite already being trained on a large text dataset, GPT2 and GPT3 can be further fine-tuned with text in a given domain. In addition, it is possible to set hyperparameters such as temperature and truncation token. For our purpose, we fine-tuned the models with our collection of documents

System	$\overline{BLEU}$	σ	$\overline{BERTScore}$	σ
Whoosh	74.96	±0.04	86.26	±0.05
BERT + Whoosh	88.96	±0.07	81.67	±0.05
GPT2	57.36	±0.11	88.88	±0.04
GPT3	84.97	±0.04	85.66	±0.04

Table 1: Average BLEU and BERTScore of Selected Approaches.

and empirically chose the following hyperparameters for both GPT2 and GPT3: temperature, which measures text generation randomness from 0 to 1, was set to 0.2, thus avoiding highly random text generation; the truncation token was set to '\n\n', meaning that, upon reaching this token, both models would stop generating and the answer would be ready to retrieve. The posed question was used as the prefix of the text to be generated.

The BERT+Whoosh combination used the previously created Whoosh index as starting point. As in the simple Whoosh, questions were used as search strings. However, in this case, Whoosh simply retrieves the most relevant document(s), then given to BERT as a context for extracting the answer to the question. Even though Whoosh could be used for retrieving more than one document, we decided to retrieve only the most relevant, following our experimentation, where results were similar for 1, 3 or 5 documents. The BERT-QA model is already fine-tuned for open-domain QA, and did not require any additional training.

3.3 Goal

For getting insights on necessary human labour, answer time and quality, the work was divided in two distinct phases: (i) **setup**, where systems were prepared to be used, covering any necessary configurations; (ii) **evaluation**, where each system was used for answering the questions and a set of metrics was computed from their answers. In the first phase, we subjectively assess the required human labour for putting such a system to work. In the second, we assess answer quality and answering time.

For assessing the quality of the answers, we adopted two metrics, BLEU [5] and BERTScore [8], typically used in machine translation and QA. BLEU is based on the n-gram overlap between the correct and the retrieved answer. On the other hand, BERTScore aligns both texts according to the meaning of each token, obtained from their representation in a given BERT model, to finally compute precision, recall and F1.

Time spent between submitting the documents and getting the answer for the last question was measured and then divided by the total number of questions, thus providing the average time per answer. Even if the result of the preparation could be reused, the total time included Whoosh's indexation and GPT's fine-tuning. With the exception of GPT3, ran in OpenAI's API, experimentation was conducted on a Google Colab notebook[5].

4 Evaluation

Table 1 presents the average quality scores of each approach, when used to answer the compiled questions. Presented BLEU scores are cumulative 4-gram, i.e., BLEU is computed for 1 to 4-grams with resulting scores aggregated by a geometric mean. To compute BERTScore, we relied on the pre-trained version of BERTimbau. Presented scores are the F1.

GPT3 got the third highest BERTScore and the second highest BLEU, respectively 85.66% and 84.97%. It also had the fastest answering time of 3.4 seconds, but, unlike the other approaches run on Colab, it was run through OpenAI's API. For instance, GPT2 had an average answering time 35 times greater.

GPT2 got a BLEU score of 57.36% and a BERTScore score of 88.88%. Despite having only a minor difference in BERTScore, when compared to the other models, BLEU scores were at least 15 points lower than for the other models. This is a consequence of a generating approach, which might use different tokens than expected.

Also one of the fastest to retrieve an answer, only 0.4 seconds behind GPT3, search engine Whoosh proved to have the third highest BLEU score (74.96%) and the second highest BERTScore score (86.26%), presenting a competitive performance for a baseline.

Lastly, the BERT+Whoosh combination got a BERTScore of 81.67%, slightly below the other approaches but nonetheless very similar, and the highest BLEU, 88.96%, which is not surprising, as it answers with spans of the given context. Note that, even though BERT-QA was not training for our domain, it adapted well. It also revealed an acceptable average answering time, 5 times slower than GPT3 and Whoosh, but 6 times faster than GPT2.

[5] https://colab.research.google.com/

5 Conclusion

We explored a set of approaches for QA in Portuguese, with the aim of taking conclusions on their configuration complexity, answering time, and quality when obtaining answers from a collection of technical documents. IR-based Whoosh and transformer language models BERT, GPT2 and GPT3 were configured and submitted to experimentation, where they were provided with set of documents and questions, for which they had to get the answers. The latter were assessed with BLEU and BERTScore.

Even if achieving some of the best scores in both quality metrics whilst having an easy through API configuration, GPT3's experiments were significantly limited, as it had a long access waiting period and came with a high cost associated to its usage. Similarly unreasonable is the time consumption associated with the process of fine-tuning GPT2 to the given domain, which also tends to escalate with the size of the domain.

Comparing simple Whoosh with its combination with BERT, the latter got a slightly higher average score on answer quality, but took a greater time to get the answers and a more complex configuration, as it was necessary to configure both systems and their integration. However, the major difference lies in the type of answer each approach retrieves. Whilst BERT is capable of providing an intelligible response, Whoosh simply retrieves scattered words, which may not make sense together.

Thus, considering the three assessed aspects, our choice of an approach for answering Portuguese technical documents lies on the combination of the transformer model BERT with the IR-based Whoosh. Despite having only experimented with a small collection of 25 documents and 67 questions, on a single domain, we believe that this combination could adapt well to much larger collections, possibly increasing the number of documents retrieved by Whoosh, and domains, given that BERT-QA answers questions in any domain.

Nevertheless, in the future, we aim to further explore the set of approaches with different and larger collections, as well as to implement a software package containing all the approaches ready to be used, given a domain and a set of questions.

Acknowledgements This work is funded by the project POWER (grant number POCI-01-0247-FEDER-070365), co-financed by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020), and by the Competitiveness and Internationalization Operational Programme (COMPETE 2020).

References

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv:2005.14165 preprint*, 2020.

[2] Paweł Budzianowski and Ivan Vulić. Hello, it's GPT-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc 2019 Conf of North American Chapter of the ACL: Human Language Technologies*, pages 4171–4186. ACL, 2019.

[4] Oleksandr Kolomiyets and Marie-Francine Moens. A Survey on Question Answering Technology from an Information Retrieval Perspective. *Information Sciences*, 181(24):5412–5434, Dec 2011.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc of 40th annual meeting of the ACL*, pages 311–318, 2002.

[6] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[7] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Proc of Brazilian Conf on Intelligent Systems (BRACIS 2020)*, volume 12319 of *LNCS*, pages 403–417. Springer, 2020.

[8] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*, 2019.

Sentinel 2 Image Scene Classification: A Comparison Between Bands and Spectral Indices

Kashyap Raiyani[1,3]
kshyp@uevora.pt
Teresa Gonçalves[1,3]
tcg@uevora.pt
Luis Rato[1,2,3]
lmr@uevora.pt

[1] Departamento de Informática, Universidade de Évora, Portugal
[2] CIMA, Universidade de Évora, Portugal
[3] EaRSLab, Universidade de Évora, Portugal

Abstract

Given the continuous increase in the global population, the food manufacturers are advocated to either intensify the use of cropland or expand the farmland, making land cover and land usage dynamics mapping vital in the area of remote sensing. In this regard, identifying and classifying a high-resolution satellite imagery scene is a prime challenge. Several approaches have been proposed either by using static rule-based thresholds (with limitation of diversity) or neural network (with data-dependent limitations). This paper adopts an inductive approach to build classifiers from spectral reflectances, comparing usefulness of the various spectral indices to raw bands information. More specifically, it considers Sentinel 2 data for six classes Scene Classification (Water, Shadow, Cirrus, Cloud, Snow and Other). The experimental results show that using raw bands performs equally well, claiming that raw bands information can be used as a replacement of the spectral indices.

1 Introduction

The integrated use of satellite and ground-based observations is widely recognized as the most feasible approach for the measurement and long-term monitoring of terrestrial variables needed by scientific investigators and decision-makers around the world. In particular, Earth observation applications are making use of the unique, synoptic capabilities of an ever-increasing number of satellite remote sensing imaging systems. A key challenge is to ensure that such measurements yield self-consistent and accurate geophysical and biophysical data over time and space, even though the measurements are made with a variety of different sensors under different observational conditions. In such Earth observations techniques, optical satellites play a major role, and one such satellite is Sentinel-2. Sentinel-2 is part of the Earth observation mission from the Copernicus Programme and systematically acquires optical imagery at a high spatial resolution over land and water bodies using 13 sensors (also known as bands). The band value (also known as surface reflectance) is defined as the fraction of incoming solar radiation reflected from Earth's surface for a specific incident or viewing case.

According to a study by the International Centre for Integrated Mountain Development (ICIMOD) [9], band ratios are used to remove undesirable effects on recorded radiances (e.g. variable illumination) since topographic slope and aspect, shadows or seasonal changes can cause differences in brightness values between identical surface materials. As a result, the interpreter's ability to correctly identify surface material in an image is hampered. The band ratio transformations can be used to mitigate these effects. Aside from that, the Spectral Indices can be used to model, predict, and track land change processes.

Between the last two decades (1999–2009 and 2009–2019), Polykretis *et al* [6] examined the impact of various spectral indices in detecting land cover changes on the Greek island of Crete. To do so, five index combinations were provided, resulting in a kappa index of 0.60–0.69 and overall accuracy of 0.86–0.96. According to Dixit *et al* [2], the visible, NIR, and SWIR bands are the most commonly used reflectance and absorptive properties for developing snow/ice cover mapping; based on these, they proposed the Snow Water Index (SWI) with an overall accuracy of 0.93 and kappa statistics of 0.94. Separately, according to Zhai *et al* [1], the majority of existing cloud/shadow detection methods are based on visible and infrared spectral band configurations with working mechanisms relatively complex and computationally complicated; as such, they proposed an unified cloud/shadow detection algorithm based on spectral indices with a cloud detection accuracy of 0.98 and a cloud shadow detection accuracy of 0.84.

Referring to all the previous work and approaches, this paper reports simulation study and encapsulating the below-mentioned contributions:

1. Focusing on the problem of optical satellite image scene classification;
2. Classification using property-specific spectral indexes and Sentinel-2 raw bands;
3. A comparison of results concerning the feature space and classification accuracy;
4. Claiming "raw bands do have impacts over overall classification" and can be used as a replacement to the spectral indices.

The rest of the document is structured as follows: Section 2 details the dataset acquisition and spectral indexes used for the study, Section 3 presents the experimental setup and obtained results and Section 4 summarises the findings and states future steps.

2 Materials

2.1 Dataset Acquisition

Raiyaini *et al* [7] published an extended database of manually labeled Sentinel-2 spectra with 13 bands values. The database consists of images acquired over the entire globe and comprises 6.6 million points (exactly 6,628,478 points) classified into one of the six classes (Water, Shadow, Cirrus, Cloud, Snow, Other). Table 1 describes the database.

Header	Description
Product ID	78 character string
Coordinates	Latitude and longitude
Bands/Features	Band 1 to 12 and 8A (13 values)
Scene	Class (Water, Shadow, Cirrus, Cloud, Snow, Other)

Table 1: Extended Sentinel-2 Database with Surface Reflectances [7].

Besides the dataset, 3 different classifier models were also published using Random Forest (RF), Extra Trees (ET) and Decision Trees (DT) (trained over default values) along with the corresponding training (50 products/images with 5,716,330 observations) and test sets (10 images with 912,148 observations). The micro-F1 [5] performance measured over the test set is presented in Table 2.

Class	RF	ET	DT	Support
Water	0.90	0.90	0.83	117010
Shadow	0.80	0.81	0.67	155715
Cloud	0.82	0.81	0.72	134315
Cirrus	0.71	0.73	0.59	175988
Snow	0.88	0.88	0.82	154751
Other	0.80	0.83	0.71	174369
micro-F1	0.81	0.82	0.72	912148

Table 2: micro-F1 performance using Sentinel-2 13 Bands values.

2.2 Spectral Indices

Spectral indices are combinations of the pixel values from two or more spectral bands in a multispectral image. Spectral indices are designed to highlight pixels showing the relative abundance or lack of a land-cover type of interest in an image. From the indexes available in the literature, a

subset was chosen specifically to identify each of the specific six classes. These are enlisted in Table 3.

Class	Spectral Indices	Reference
Water	Normalized Difference Water Index (NDWI) Sentinel-2 Water Index (SWI)	[3]
Shadow	Shadow Enhancement Index (SEI) Saturation Value Different Index (SVDI)	[8]
Cloud	Cloud Index (CI) Brightness Index (BI)	[1] [4]
Cirrus	Band 10	Sentinel-2
Snow	Normalized Difference Snow Index (NDSI) Normalized Difference Snow Ice Index (NDSII) S3 Snow Water Index (SWI)	[2]
Other	Bare Soil Index (BSI)	[6]

Table 3: Classes and Spectral Indices.

3 Experimental setup & Results

All the 3 classifiers RF, ET, and DT were trained using the default algorithm parameters parameters. The evaluation of these models (using the information from Table 3, namely 11 Indexes plus Band 10) was done using micro-F1 (F1) score [5].

Table 4 presents the results using only the spectral indices and them along with the 13 bands information. All the 3 models give similar results, having higher F1 values for Water (90%) and lower F1 values for Shadow (75%). Adding the 13 bands values to the spectral indexes does not seem to improve the results. Moreover, by comparing the results obtained using the Bands (Table 2) and the Spectral Indices (Table 4) it is possible to conclude that the use of indexes does not improve the classifier.

	Spectral Indices			13 Bands + Spectral Indices		
Class	**RF**	**ET**	**DT**	**RF**	**ET**	**DT**
Water	0.90	0.90	0.82	0.89	0.89	0.81
Shadow	0.75	0.75	0.62	0.76	0.76	0.66
Cloud	0.81	0.82	0.70	0.80	0.81	0.71
Cirrus	0.78	0.79	0.63	0.74	0.76	0.62
Snow	0.87	0.87	0.81	0.88	0.87	0.83
Other	0.79	0.80	0.66	0.81	0.81	0.71
micro-F1	0.81	0.82	0.70	0.81	0.81	0.72

Table 4: micro-F1 with Spectral Indices and 13 Bands.

By analyzing the indexes presented in Table 3, one notices that the Spectral indices models only use information from 10 bands (not included bands - 6/7/8A) of the available 13 bands of Sentinel-2. Having this in mind, classifiers were built using raw information from those 10 bands only. The obtained results are presented in Table 5 and show that there is no significant difference on classifiers performance (1% more for Random Forest and Extra Tress). Thus, we can definitively conclude that there is no need to calculate and use spectral indices instead of raw bands for Sentinel 2 Image Scene Classification (at least for studied six classes: Water, Shadow, Cirrus, Cloud, Snow and Other.)

Class	RF	ET	DT
Water	0.89	0.89	0.81
Shadow	0.76	0.76	0.66
Cloud	0.80	0.81	0.71
Cirrus	0.74	0.76	0.62
Snow	0.88	0.87	0.83
Other	0.81	0.81	0.71
micro-F1	0.81	0.81	0.72

Table 5: micro-F1 using 10 Bands (not included bands - 6/7/8A).

To further investigate, models were built using raw data only from the 10 bands used by the 12 spectral indices referred in Table 3. The obtained results, presented in Table 5, show that the performance is not decreased when compared to the models that use all 13 bands. Furthermore, looking at individual F1 values, one can see that the maximum difference of model performance over any two classes is 15%, 13%, and 21% for RF, ET, and DT respectively. (For example, with RF, Cirrus has a "worst" micro-F1 of 0.74% and Water has a "best" micro-F1 of 0.89%, generating the maximum model performance difference of 15%). Comparing to Table 4 (Spectral Indices) results, where the maximum difference is 15%, 15%, and 20% for RF, ET, and DT respectively. Apart from this, for a single class (Water and Cirrus), both the approaches (raw bands and Spectral Indices) have the maximum F1-score around 89%-90% with minimum value around 62%-63%; showcasing an equivalent performance.

4 Conclusion

Through our experiments, we were able to provide a study that proves that raw bands of Sentinel-2 can be used as features instead of using different Spectral Indices. This can be verified from the results presented on Tables 4 and 5. Moreover, when bands and spectral indices are used together no improvement is verified (Table 4).

As future work, the authors of the paper would like to incorporate radar information from Sentinel 1 and verify their impact over Water, Shadow, Cirrus, Cloud, and Snow detection.

Funding

This work was supported by the NIIAA (Núcleo de Investigação em Inteligência Artificial em Agricultura) project, Alentejo 2020 program (reference ALT20-03-0247-FEDER-036981).

References

[1] Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 144:235–253, 2018. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2018.07.006.

[2] Abhilasha Dixit, Ajanta Goswami, and Sanjay Jain. Development and evaluation of a new "snow water index (swi)" for accurate snow cover delineation. *Remote Sensing*, 11(23), 2019.

[3] Wei Jiang, Yuan Ni, Zhiguo Pang, Xiaotao Li, Hongrun Ju, Guojin He, Juan Lv, Kun Yang, June Fu, and Xiangdong Qin. An effective water body extraction method with new water index for sentinel-2 imagery. *Water*, 13(12), 2021.

[4] Richard J Kauth and GS Thomas. The tasselled cap–a graphic description of the spectral-temporal development of agricultural crops as seen by landsat. In *LARS symposia*, page 159, 1976.

[5] Juri Opitz and Sebastian Burst. Macro F1 and macro F1. *CoRR*, abs/1911.03347, 2019. URL `http://arxiv.org/abs/1911.03347`.

[6] Christos Polykretis, Manolis G. Grillakis, and Dimitrios D. Alexakis. Exploring the impact of various spectral indices on land cover change detection using change vector analysis: A case study of crete island, greece. *Remote Sensing*, 12(2), 2020.

[7] Kashyap Raiyani, Teresa Gonçalves, Luís Rato, Pedro Salgueiro, and José R. Marques da Silva. Sentinel-2 image scene classification: A comparison between sen2cor and a machine learning approach. *Remote Sensing*, 13(2), 2021. ISSN 2072-4292. doi: 10.3390/rs13020300. URL `https://www.mdpi.com/2072-4292/13/2/300`.

[8] Genyun Sun, Hui Huang, Qihao Weng, Aizhu Zhang, Xiuping Jia, Jinchang Ren, Lin Sun, and Xiaolin Chen. Combinational shadow index for building shadow extraction in urban areas from Sentinel-2A MSI imagery. *International Journal of Applied Earth Observation and Geoinformation*, 78:53–65, June 2019. doi: 10.1016/j.jag.2019.01.012.

[9] Poonam Tripathi. Remote sensing indices and their applications, 10 March, 2020.

caPAD - A context aware model for face presentation attack detection

Pedro C. Neto[1,2]
pedro.d.carneiro@inesctec.pt
Ana F. Sequeira[1]
ana.f.sequeira@inesctec.pt
Jaime S. Cardoso[2,1]
jaime.cardoso@inesctec.pt

[1] INESC TEC
Porto, Portugal
[2] Faculty of Engineering
University of Porto, Porto, Portugal

Abstract

Presentation attacks are some of the most frequent vulnerabilities of biometric systems. To perform these attacks, the impostors attempt to bypass the biometric vision system. The human visual cortex system can leverage distinct information from the background and the main focus. However, researchers still rely on the idea that the background is, in the majority of cases, harmful to machine learning algorithms. And thus, face presentation attack detection models are trained with tight crops of the face. It is argued that it rather limits the model and its performance. We further show that a binary classification system aware of the background is capable of outperforming its counterpart that gets no information regarding the background. The proposed methodology beats current approaches and achieves an equal error rate (EER) of just 0.9%. We further analyze the predictions from an interpretability point-of-view and argue that the background elements used by the model are similar to the ones used by humans.

1 Introduction

There has been an unintentional limitation to the capabilities of face presentation attack detection (PAD) systems over the years. The input information that these systems receive is limited. Most of the biometric systems are targeted by presentation attacks. The methods used to defend these systems from such attacks rely on tight face crops [10]. This means that besides the face, all the other information (background) is removed. It brings some advantages to biometric systems, for instance, on face recognition several independent faces can be processed independently. On the other hand, it removes contextual and spatial information that might be useful for the defence against presentation attacks. The human visual cortex can process this spatial and contextual information to identify some attacks on the human eye. It is even possible to verify that some replay attacks can fool humans if the replay device has a high resolution. We argue that machine vision systems can likely learn to leverage the extra information when it is available. They can even decide if the background information is useful for the prediction, or not. Hence, we believe that instead of limiting the information given to the model, researchers must aim to develop novel and robust models that are capable of leveraging contextual information. Even if it remains in a more philosophical domain, we propose that researchers goal should be to approximate the models to the human vision, or even surpass it.

Deep neural networks learn, sometimes, undesired patterns which are used for predictions. And thus, through the use of explainable artificial intelligence (Xai) methods, a qualitative assessment of the spatial information used by the model to predict if an image represents an attack was conducted. Intended to avoid errors due to opaqueness, we used visualization methods such as Grad-CAM++ [4]. Their output indicated the spatial areas used by the model. It was also verified that they correlate with the ones used by humans. So we speculate about the influence of the background in the future of face PAD algorithms.

The ROSE-Youtu dataset [12], differently from the majority of other datasets, includes a high diversity of attacks, which include both two-dimensional and three-dimensional information. Hence, it was used to study the impact of the background in the model performance and whether the background affects the capability of generalizing between attacks.

2 Dataset

The dataset selected to conduct the experiments of this paper is the ROSE-Youtu dataset [12]. It contains, in its public version, 3350 videos with 20 different subjects. The average clips length is 10 seconds. These clips were collected from five mobile devices (distinct camera resolutions for all of them) and five lighting conditions. The front-facing camera was used with a distance between face and camera of about 30 to 50 centimetres.

Table 1: List of attacks present in the ROSE-YOUTU dataset [12].

Attack	Description
-	Genuine (bona fide)
#1	Still printed paper
#2	Quivering printed paper
#3	Video which records a Lenovo LCD display
#4	Video which records a Mac LCD display
#5	Paper mask with two eyes and mouth cropped out
#6	Paper mask without cropping
#7	Paper mask with the upper part cut in the middle

Each subject is associated with eight distinct types of videos. Each type corresponds to a label. The first label, 0, represents genuine samples, whereas the remaining seven represent one of seven types of attacks. The first attacks are print attacks, whereas the third and fourth are replay attacks. The remaining are attacks based on paper masks. These attacks are described in Table 1.

3 Methodology

Despite the distinct types of attacks that endanger biometric systems, in practice, it is only necessary to infer a given image is from an impostor or a genuine person. Therefore, the face PAD problem is, in its essence, formulated as a binary classification task. To tackle this problem we trained a MobileNet v2. This backbone network is optimized to minimize the probability of wrong classes and maximize the probability of the correct class. The outputs of the network (2 values) are activated with the softmax nonlinearity. Weight optimization is done using the binary cross-entropy loss (Eq. 1).

$$BCE(y,p) = -(y\log(p) + (1-y)\log(1-p)) \quad (1)$$

4 Results & Discussion

Table 2: Comparison of the binary classification system with their versions with and without background. Attack Presentation Classification Error Rate (APCER), Bona Fide Classification Error Rate (BPCER) and Equal Error Rate (ERR) are displayed as %, and lower values are better. In bold is the best result per column.

Method	Background	APCER	BPCER	EER
Binary Classification	No	0.493	**2.199**	1.319
	Yes	**0.123**	3.051	**0.935**

The initial experiments produced intended to evaluate the effect of the inclusion and exclusion of contextual background. As expected, the performance increased dramatically when the background was available. The background provides more information, and thus all the metrics benefited from major improvements. The improvements on the equal error rate are 29%. The results can be seen in Table 2.

When compared to the state-of-the-art (Table 3), the results are even more impressive. The inclusion of background boosts the performance to be better than the performance of the other published methods.

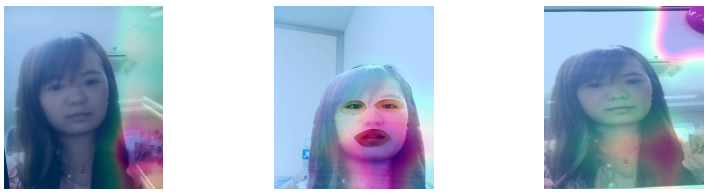
(a) Crop Attack #4 (b) Crop Attack #1 (c) Crop Attack #6 (d) Crop Genuine (e) Crop Genuine (f) Crop Genuine

(g) Attack #4 (h) Attack #1 (i) Attack #6 (j) Genuine (k) Genuine (l) Genuine

Figure 1: Samples collected from the ROSE-YOUTU dataset [12] containing images from attacks and genuine captures. On the top row, cropped images are displayed. Whereas the bottom row contains the exact same images, but with all the background information included.

Table 3: Comparison of the proposed approach with the state-of-the-art. EER is displayed as %. In bold is the best result per column.

Method	EER
Color LBP [1, 5]	27.6
CoALBP (YCBCR) [12]	17.1
CoALBP (HSV) [12]	16.4
Color [2, 5]	13.9
De Spoofing [5, 9]	12.3
RCTR-all spaces [5]	10.7
ResNet-18 [7]	9.3
SE-ResNet18 [8]	8.6
AlexNet [12]	8.0
DR-UDA (SE-ResNet18) [13]	8.0
DR-UDA (ResNet-18) [13]	7.2
3D-CNN [11]	7.0
Blink-CNN [6]	4.6
DRL-FAS [3]	1.8
Ours	**0.9**



(a) Replay Attack (b) Paper Mask Attack (c) Print Attack

Figure 2: Explanations produced for a prediction from a frame from a video of subject #23. Colors closer to pink represent areas with larger relevance for the decision. Bluish colors represent less important pixels.

Finally, we produced explanations of our model for an example of each category of attacks. For the replay attack, we produced the explanations in Figure 2a. That figure shows that the model leveraged the presence of reflections in the attack image. Figure 2b shows the explanation for a paper mask attack, and as expected, the explanation does not rely on the background. Instead, the model directs its focus to the mask area for the final prediction. Finally, the print attack explanation is seen in Figure 2c. It shows that the model understands the conditions of the image given and directs its focus to an important background artefact, the pin holding the image.

5 Conclusions

In this work, we explored our belief that researchers have been creating limitations for face presentation attack detection models by cropping the face from the frame. The experiments of our work corroborated the view that a face PAD model is capable of leveraging both background and face elements to make a correct prediction.

This proposed approach surpassed the state-of-the-art results for the ROSE-YOUTU dataset. The lightweight model is capable of providing impressive results. The interpretability analysis corroborated the beliefs regarding the usage of background elements.

Acknowledgements This work was partially funded by the Project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program - NORTE 2020, by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership, and within the PhD grant "2021.06872.BD"'.

References

[1] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *2015 IEEE international conference on image processing (ICIP)*, pages 2636–2640. IEEE, 2015.

[2] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, 2016. doi: 10.1109/TIFS.2016.2555286.

[3] Rizhao Cai, Haoliang Li, Shiqi Wang, Changsheng Chen, and Alex C Kot. Drl-fas: A novel framework based on deep reinforcement learning for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 16:937–951, 2020.

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. doi: 10.1109/WACV.2018.00097.

[5] Yuting Du, Tong Qiao, Ming Xu, and Ning Zheng. Towards Face Presentation Attack Detection Based on Residual Color Texture Representation. *Security and Communication Networks*, 2021:6652727, 2021. ISSN 1939-0114. doi: 10.1155/2021/6652727. URL https://doi.org/10.1155/2021/6652727.

[6] Md. Mehedi Hasan, Md. Salah Uddin Yusuf, Tanbin Islam Rohan, and Shidhartho Roy. Efficient two stage approach to detect face liveness : Motion based and deep learning based. In *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, pages 1–6, 2019. doi: 10.1109/EICT48899.2019.9068813.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[9] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 290–306, 2018.

[10] Dakshina Ranjan Kisku and Rinku Datta Rakshit. Face spoofing and counter-spoofing: a survey of state-of-the-art algorithms. *Transactions on Machine Learning and Artificial Intelligence*, 5(2):31, 2017.

[11] Haoliang Li, Peisong He, Shiqi Wang, Anderson Rocha, Xinghao Jiang, and Alex C. Kot. Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(10):2639–2652, 2018. doi: 10.1109/TIFS.2018.2825949.

[12] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C. Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809, 2018. doi: 10.1109/TIFS.2018.2801312.

[13] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 16:56–69, 2021. doi: 10.1109/TIFS.2020.3002390.

Predicting soil electro-conductivity using Sentinel-1 images

Eduardo Medeiros[1]
efarofia@uevora.pt
Sajib Ahmed[1,3]
sajib@uevora.pt
Teresa Gonçalves[1,3]
tcg@uevora.pt
Luís Rato[1,2,3]
lmr@uevora.pt

[1] Departamento de Informática, Universidade de Évora, Portugal
[2] CIMA, Universidade de Évora, Portugal
[3] EaRSLab, Universidade de Évora, Portugal

Abstract

The quality and yield of a soil can be measured by using a wide range of soil indicators. One such indicator is soil's electro-conductivity which is an excellent indicator of the presence of soil nutrients. This work aims to create a machine learning model to predict the soil's electro-conductivity (EC) using radar images from the satellite Sentinel-1. Using EC readings from 14 corn field parcels and Sentinel-1 readings over the course of one agriculture year, several regression models were generated. These models were designed using information from the full agriculture year or only 3 months, both or only one of the VV and VH polarisations. The results show that when using a full year data VV and VH polarisations are able to generate models with similar performance (R^2 of 0.888 for VH and 0.884 for VV) but when using only 3 months data, only April to June trimester using both polarisations are able to reach similar a performance (R^2 of 0.867); moreover VH polarisation seems to carry out more descriptive information when compared with VV (specially when using only 3 months data). Finally, performance results seem to be independent of the yearly radar data time-window.

Keywords: Soil electro-conductivity, Remote sensing, Sentinel-1, Regression, K-nearest neighbours

1 Introduction

Precision farming incorporates a series of strategies and tools that allow farmers to optimise and increase soil quality and productivity putting in place a set of targeted key interventions. These interventions are selected based on collected information of minerals, nutrients, water, soil texture, drainage conditions, salinity, and other soil characteristics over farmland [3]. Soil electro-conductivity (EC) is one of simplest, and least expensive soil measurements available to precision farming [5].

Recently, soil properties are being obtained using remote sensing techniques [2]. Sentinel-1 [4] is a synthetic aperture radar instrument (SAR) satellite that consists of a set of two satellites, Sentinel-1A and Sentinel-1B, which share the same orbital plane with a 12-day revisiting period. This set of satellites provides images in two different polarisations: VV (vertical transmit, vertical receive) and VH (vertical transmit, horizontal receive).

In a previous work [1], Sentinel-1 information was used to build models able to classify soils as sandy, free and clayish (by discretizing EC values) achieving F1-scores between 54.44% and 75.6% for clayish and free soils, respectively, over a test set of 13001 points. The current work, instead, aims at predicting the EC value itself. Besides studying if polarisations have different discriminative power, it also studies which months are more informative and if the sentinel data from different years generates models with similar performances.

The rest of the paper is organized as follows: Section 2 introduces the data sets and algorithms used, describes the experiments performed and the experimental setup; Section 3 presents and discusses the results obtained; finally, Section 4 concludes the paper and presents future work.

2 Materials and Methods

2.1 Data sets

The on-site EC values were obtained between March 28 and May 3, 2016, on a set of 14 parcels of corn fields located in Alentejo with a 10-meter interval resulting in a total of 65003 points. Measured values ranged between 0.226 and 240.592; Figure 1 presents the distribution of the values.

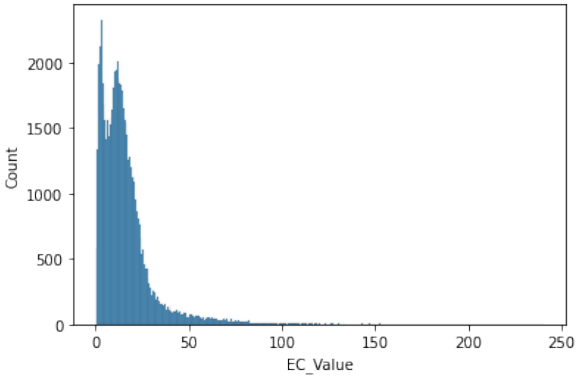


Figure 1: Histogram of EC values on the data set.

Radar data was collected from two time windows each corresponding to full agricultural years: (a) from October 6, 2018 to September 25, 2019 (the most recent year available when data was collected) and (b) from October 3, 2015 to September 28, 2016 (the year corresponding to the EC readings). The data extracted for the 2018-2019 time window corresponds to orbit 147 from both satellites because for Sentinel-1B there was only data for that orbit (although information from more orbits existed for Sentinel-1A). For 2015-2016 time window data was collected from 3 available orbits: 43, 50 and 145.

Similarly to the previous study [1], the 2018-2019 data is composed by 120 descriptive attributes corresponding to VV and VH values from both satellites over the considered time window (60 different dates; 30 from each satellite). Since satellite Sentinel-1B was only launched in 2016, the 2015-2016 data is composed by VV and VH values from Sentinel-1A, composed by 27, 26 and 29 different dates for orbits 43, 50 and 145, respectively.

Table 1 presents a characterisation of the radar values for each polarisation and satellite during the 2018-2019 time window. It is easly seen that the range of values for VH polarisation is much smaller than the range for VV polarisation. On the other hand, both satellite present similar ranges.

	Sat A		Sat B		Sat A+B	
	VH	VV	VH	VV	VH	VV
mean	95.81	216.82	88.09	219.49	91.95	218.16
std	24.93	66.60	24.05	67.88	24.80	67.26
min	28.33	62.11	25.11	58.78	25.11	58.78
25%	77.78	173.56	70.89	173.44	74.11	173.56
50%	94.78	207.89	86.78	209.89	90.67	208.89
75%	111.89	248.44	103.00	253.78	107.67	251.11
max	250.44	2009.33	242.22	1988.78	250.44	2009.33

Table 1: Characterisation of the attributes over a one year period.

2.2 Algorithms

Several machine learning algorithms for regression were tested namely Support Vector Machines (SVM), K Nearest-Neighbours (KNN), Ridge and Lasso.

2.3 Experiments

A first set of experiments was done using the 2018-2019 data to study the algorithms and the sets of features that generate the most performing

models. SVM was tested with linear kernel and $C = \{0.1, 1, 10\}$, KNN tested with $K = \{1, 5, 9\}$, Ridge with $\alpha = \{0.1, 1, 10\}$ and Lasso with $\alpha = \{1, 0.01, 0.0001\}$ (the rest of the parameters were the default for all algorithms). Models were built using VV and VH values, alone and together, for all available dates (60 attributes per polarisation), by monthly averaging them (12 attributes per polarisation) and by using trimester values (15 attributes per polarisation) instead of the full year.

A second set of experiments, using the 2015-2016 data (using one and both polarisations), was also experimented, aiming to check if the models performed differently using radar information from different years.

2.4 Experimental Setup

For developing the models Python (v3.7.9) with scikit-learn (v0.23.2) were used and a stratified train-test split generated with 75% for training (48752 samples) and 25% for testing (16250 samples). The models were evaluated using the coefficient of determination, R^2, a performance measure that normally ranges between 0 and 1, with 0 corresponding to a constant model that always predict the training test average value and 1 corresponding to a perfect prediction. Experiments were also performed to check if the existence of outliers (namely, very high values of VV) influenced the results; no influence was found.

3 Results

This section presents and discusses the results obtained with 2018-2019 and 2015-2016 time windows data sets.

3.1 2018-2019 data

First, a set of experiments, aiming to find the algorithm and parameters, was performed using both polarisations and all dates (120 attributes) and a 5-folds cross-validation over the training set. Table 2 presents the results over the test set for the best parameters found. As can be seen, KNN performs best by a large margin.

Algorithm	Parameters	R^2
LinearSVM	$C = 1$	0.243
KNN	$K = 5$	**0.883**
Ridge	$\alpha = 1$	0.331
Lasso	$\alpha = 0.01$	0.331

Table 2: R^2 results for the best performing parameters.

Then, a more thorough search over the KNN parameters was conducted (also using a 5-folds cross-validation over the training set) with K={1,3,5,7,9}, weights={*uniform*, *distance*} and Minkowski distance with $p = \{1, 2, 3\}$. The parameters with best results were $K = 3$, $p = 1$ and weights=*distance*.

Finally, using the fine-tuned parameters, the different sets of attributes mentioned in subsection 2.3 were tested. Table 3 presents results using the full year data (all dates and monthly average) and trimester data.

Pol	all	m. avg	Oct-Dec	Jan-Mar	Apr-Jun	Jul-Sep
VH	**0.888**	0.777	0.636	0.662	0.777	0.718
VV	0.884	0.716	0.606	0.571	0.714	0.657
both	0.886	**0.860**	0.839	0.848	**0.867**	0.854

Table 3: R^2 results for full year and trimester dates.

Looking at the full year data one can conclude that, when using yearly individual dates, the results obtained are very similar using one (60 attributes) or both polarisations (120 attributes), with VH presenting the best result with $R^2 = 0.888$. This is not true when using monthly values: the best results, by a large margin, are obtained using information from both polarisations (12 attributes for each polarisation), with a result of $R^2 = 0.860$. Nonetheless, when comparing single polarisations, VH continues to carry more discriminant information than VV ($R^2 = 0.777$ vs. $R^2 = 0.716$).

When comparing trimester data one can conclude that Apr-Jun trimester data is the most informative one reaching a value of $R^2 = 0.867$ with both polarisations (30 attributes). On the other end Oct-Dec trimester data is the least informative. Also, when using trimester data, using both polarisations increase the performance substantially (more than 10%, with higher increases for the least performing trimesters).

3.2 2015-2016 data

As previously mentioned, this set of experiments aimed at checking if the models performed differently using radar information from different years; 2015-2016 time window was chosen to include the on-site EC readings. Since Sentinel-1B was only launched in 2016, this data only contains values from Sentinel-1A, but readings of the 3 available orbits were collected. Parameter fine-tuning was also conducted over KNN algorithm, with the best performance obtained with the the same parameters.

In order to compare the results between the two time windows, an additional experiment with 2018-2019 data was conducted using only the Sentinel-1A readings. Table 4 presents the results.

	2015-2016			2018-2019
Pol	Orb 43	Orb 50	Orb 145	Orb 147
VH	0.869	0.869	0.880	0.876
VV	0.865	0.879	0.876	0.873
both	0.880	0.876	0.885	0.886

Table 4: 2015 data set scores for each orbit and polarisation combination.

As can be seen from the table, when using a full year radar information, different orbits generate models with similar performances (with higher values for orbit 145 possibly because it contains more dates); once again VH polarisation outperforms VV and a minor improvement over VH performance is observed when adding VV information. On the other hand, it is easily seen that the results seem to be independent of the time window of the collected radar data.

4 Conclusions and Future Work

This work presents a regression model to determine the soil eletro-condutivity using radar information. The developed models reached a R^2 performance of 88.8% using data from both satellites and a full year time window (2018-2019 data). Nonetheless, similar performances were obtained using information from just one satellite (88.6%). Moreover, VH polarisation seems to carry out more descriptive information when compared with VV, obtaining a similar performance to using both polarisations (when using a full year time window). Also, when using trimester data, Apr-Jun has the highest R^2 (86.7%) while Oct-Dez has the lowest (83.9%). Finally, performance results seem to be independent of the yearly radar data time-window.

As future work, we intend further investigate the use of radar data aiming to produce better models (joining info from different orbits and/or different years) and to build an application that, given a site's set of radar images, is able to generate the corresponding electro-conductivity map.

Funding

This work was supported by NIIAA (Núcleo de Investigação em Inteligência Artificial em Agricultura) project, Alentejo 2020 program (reference ALT20-03-0247-FEDER-036981).

References

[1] Sajib Ahmed, Teresa Gonçalves, Luís Rato, Pedro Salgueiro, J. R. Marques da Silva, and Luis Paixão Filipe Vieira. Classifying soil type using radar satellite images. 2020.

[2] Yufeng Ge, J Alex Thomasson, and Ruixiu Sui. Remote sensing of soil properties in precision agriculture: A review. *Frontiers of Earth Science*, 5(3):229–238, 2011.

[3] Robert Dwight Grisso, Marcus M Alley, David Lee Holshouser, and Wade Everett Thomason. Precision farming tools. soil electrical conductivity. -, 2005.

[4] Paul Snoeij, Evert Attema, Malcolm Davidson, Berthyl Duesmann, Nicolas Floury, Guido Levrini, Björn Rommen, and Betlem Rosich. The sentinel-1 radar mission: Status and performance. In *2009 International Radar Conference" Surveillance for a Safer World"(RADAR 2009)*, pages 1–6. IEEE, 2009.

[5] Iulian-Florin Voicea, Mihai Matache, and Valentin Vladut. Researches regarding the electro-conductivity determination on different soil textures from romania, before sowing. *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Agriculture*, 66(1), 2009.

Complementary and case-based explanations for clinical decision support

Wilson Silva
wilson.j.silva@inesctec.pt
Jaime S. Cardoso
jaime.cardoso@inesctec.pt

INESC TEC and University of Porto
Porto, Portugal

Abstract

Currently, clinicians and radiologists face an excessive workload due to the large number of medical images they have to analyse, which may increase undesired diagnosis mistakes. In the last years, several deep learning methods were proposed in order to act as decision support systems, easing the burden posed upon clinicians. However, deep learning methods are highly complex and lack interpretability. Thus, interpretability and explanatory evidence are of utmost importance, contributing both to increasing trust and as decision support. This work aims to be a step towards the clinical integration of complex models and the explanations produced through them. By exploring the usage of complementary and case-based explanations, we are able to convince and help clinicians in their diagnosis process.

1 Introduction

The use of deep learning algorithms in the clinical context is hindered by their lack of interpretability. One way of increasing the acceptance of such complex algorithms is by providing explanations of the decisions. Besides helping to understand model behaviour, the explanatory process also supports the decision-making of the radiologist or clinician under challenging diagnosis scenarios. Our work investigated strategies to provide decisions, complementary and case-based explanations in several clinical applications, such as aesthetic evaluation of breast cancer treatments, melanoma detection, and pleural effusion diagnosis. This article aims to provide an overview of the work presented in [4, 5, 6], highlighting the major developments and current challenges.

2 Methods

The first method we proposed was a Deep Neural Network (DNN). This DNN had some regularization techniques to make it more interpretable, such as the use of monotonic constraints in part of the layers of the network. This DNN (Fig. 1) possessed two branches, one to process the monotonic features and another to process the non-monotonic features. After some processing layers, information was merged between the two, followed by layers with monotonic constraints, and a diagnosis decision.

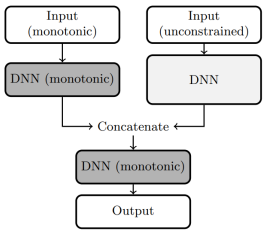


Figure 1: Deep Neural Network for Complementary Explanations [4]

Besides the classification decision, we used this network to come up with complementary explanations in the form of rule-based verbal explanations and case-based visual explanations. Rule-based explanations were extracted by estimating the local contribution of each feature, while case-based explanations were found by identifying the nearest neighbour in the semantic space previous to the classification decision (both the nearest neighbour of the same class and of the opposite class). Moreover, by performing a sensitivity analysis with regards to this semantic space, one can understand which features are making two cases similar or different, leading to additional verbal explanations accompanying the visual ones.

Pursuing even further this idea of explanation diversity, we proposed another model, an Ensemble Model [5] (Fig. 2). This ensemble was constituted by the previously proposed DNN, but also by a Scorecard (type of model commonly used in finance due to being highly interpretable) and a Random Forest (in itself an ensemble model).

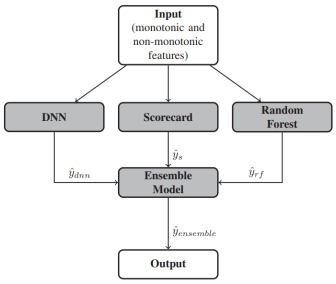


Figure 2: Ensemble Model for Complementary Explanations [5]

From this Ensemble Model, we were able to come up with a diagnosis decision (by majority voting). We could also generate verbal rule-based explanations based on the Scorecard and Random Forest and case-based explanations based on the DNN (with the same process as described before). In order to extract a single explanation for our general Ensemble Model but also for the Random Forest, we proposed the method described in Eq. (1), where the global explanation for the ensemble ($E_{global}(x)$) is the explanation produced through the model ($E_n(x)$) from the pool of M models that agrees with the ensemble decision, and that leads to the explanation with the highest correctness ($corr(E_n(x))$). In this context, N is the total number of models in the ensemble, and correctness is a measure of explanation accuracy, also proposed in one of our works [4].

$$E_{global}(x) = \underset{E_n(x)}{\text{argmax}}(corr(E_n(x)), n \in \{1,...,M\} \wedge M \leq N \qquad (1)$$

The works presented before have as input high-level features related to well-defined clinical concepts. However, sometimes those concepts are not defined, or there is no annotation available. Thus, our follow-up work was built to have images as inputs. Moreover, after discussions with clinicians, we focused our attention on case-based explanations and medical image retrieval. Looking to similar images feels more natural to radiologists than receiving a decision and a verbal explanation of how it originated. In their regular workflow, radiologists typically look for similar disease-related images when dealing with a dubious diagnosis case. Therefore, we intended to come up with a system that automated that extremely time-consuming search. Since disease information is present in just a tiny region of the image, we proposed a method that takes that information into account, trying to find the respective region in a weakly supervised manner, thus, discarding the need for additional annotations. Our method consisted of a two-step training procedure. The first step consisted of a "normal" training process of a Convolutional Neural Network (CNN). Afterwards, interpretability saliency maps were computed, and the CNN was fine-tuned with them, increasing the focus of the network in the disease-related regions of the image.

3 Results

We applied the methods from [4, 5] to two different clinical applications: Aesthetic Evaluation of Breast Cancer Treatments and Melanoma Detection. Here, we focus only on the Aesthetic Evaluation of Breast Cancer Treatments [1], where we considered 143 images, 23 high-level features, and binary classification problems (defined based on the 4-class ground-truth). As an example, results obtained with our DNN for one test image are presented in Fig. 4. For the experimental assessment, explanations

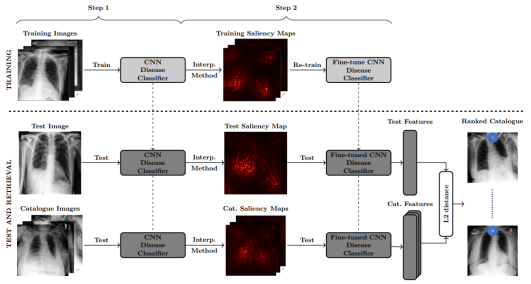


Figure 3: Interpretability-guided Content-based Medical Image Retrieval [6]

were shown to clinicians, who validated their meaningfulness. Moreover, quantitative measures of the explanations were computed based on the 3Cs of interpretability: completeness, correctness, and compactness [4].

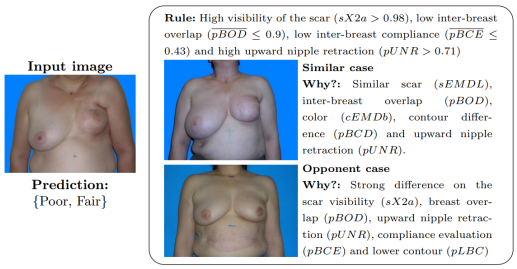


Figure 4: Complementary Explanations provided by Deep Neural Network for the Aesthetic Evaluation of Breast Cancer Treatments [4]

Regarding our medical image retrieval method, experiments were performed with the CheXpert dataset [2], a large and publicly available dataset composed of thoracic X-ray images. We extracted five different catalogues of images from this dataset, which a board-certified radiologist ranked in terms of similarity to a respective test image. In Fig. 5, we present an example of a test image, and of the top retrieved cases given by each of the methods considered (EXPERT is our board-certified radiologist, SSIM is a method based on the Structural Similarity Index, CNN is the model resultant from step 1 of our model, and IG-CBIR is the model resultant from step 2).

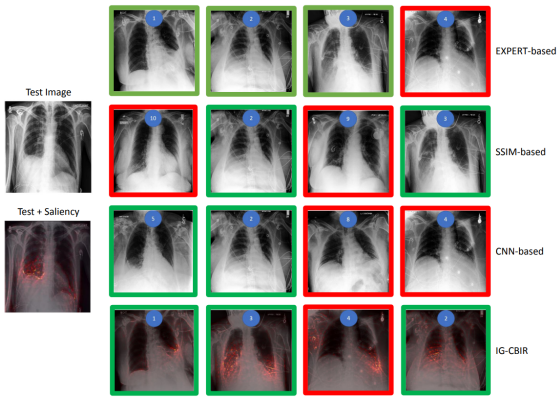


Figure 5: Retrieved catalogue images for one example test image. [6]

Furthermore, the methods were also evaluated quantitatively by computing the normalized Discounted Cumulative Gain (nDCG). Results are shown in Fig. 6, demonstrating that our method is the one that follows more closely the rationale of a radiologist.

4 Conclusions and Future Work

This work follows the premise that machine learning explanations can be as important as machine learning decisions supporting the clinical diagnosis. We started with clinical applications where high-level clinical concepts are well-defined, building models that process that information to make a diagnosis decision. Moreover, the models were built to allow the generation of complementary explanations, consisting of rule-based

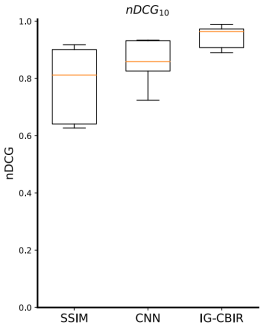


Figure 6: Box-and-whisker plot regarding the nDCG results for top-10 retrieved images. [6]

verbal explanations and case-based visual explanations. We further investigated the domain of case-based explainability through the retrieval of similar disease-related well-curated cases. Considering medical image retrieval, it became clear that general image retrieval techniques were not suitable, as relevant clinical information is localized in a very specific region of the image. In order to provide additional focus on the representations, we came up with a novel method based on interpretability techniques. Even though the retrieval results obtained were promising, this case-based retrieval approach may not find a way into clinical workflow, as the person's identity in the explanatory image is exposed. To overcome that issue, we started looking for privacy-preserving methods in the literature, exploring the use of one of the most promising deep learning methods (PPRL-VGAN) for the anonymization of case-based explanations [3]. Further work is required to deploy such a system in the clinics, namely, developing better anonymization methods and integrating causal knowledge into the approach.

Acknowledgements

This work was partially funded by the Project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership, and also by the Portuguese Foundation for Science and Technology - FCT within PhD grant number SFRH/BD/139468/2018.

References

[1] Jaime S Cardoso and Maria J Cardoso. Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. *Artificial intelligence in medicine*, 40(2):115–126, 2007.

[2] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[3] Helena Montenegro, Wilson Silva, and Jaime S Cardoso. Towards privacy-preserving explanations in medical image analysis. *Workshop on Interpretable Machine Learning in Healthcare at the International Conference on Machine Learning (ICML)*, 2021.

[4] Wilson Silva, Kelwin Fernandes, Maria J Cardoso, and Jaime S Cardoso. Towards complementary explanations using deep neural networks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 133–140. Springer, 2018.

[5] Wilson Silva, Kelwin Fernandes, and Jaime S Cardoso. How to produce complementary explanations using an ensemble model. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[6] Wilson Silva, Alexander Poellinger, Jaime S Cardoso, and Mauricio Reyes. Interpretability-guided content-based medical image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–314. Springer, 2020.

Real-Time Head Movement Analysis in Teleconsultation for Depressive Disorder Assessment

Diogo Ramalho[1]
diogo.ramalho1998@gmail.com
Vasco Duarte[1]
vasco.teodoro@tecnico.ulisboa.pt
Hugo Plácido da Silva[2]
fhsilva@lx.it.pt
Miguel Constante (PhD, MD)[3]
miguel.constante@hbeatrizangelo.pt
João Sanches[1]
jmrs@tecnico.ulisboa.pt

[1] Institute for Systems and Robotics (ISR), LARSyS
Instituto Superior Técnico, Department of Bioengineering,
University of Lisbon, Lisbon, Portugal

[2] IT - Instituto de Telecomunicações
Instituto Superior Técnico
Lisbon, Portugal

[3] Department of Psychiatry, Hospital Beatriz Ângelo
Loures, Portugal

Abstract

Depression is the leading cause of inability within the active population. The way depression is currently diagnosed is based on the clinician's subjective examination of patient's mental status and thus lacks biomarkers to support medical decision. Psychomotor retardation is one of the nine symptoms that are part of DSM-V criteria for major depressive disorder (MDD) diagnosis. One of its many manifestations is reduced head movement, which has shown to be an important metric to predict depression.

In this work we present a method for patient head movement analysis during teleconsultation. The analysis results are shown in a plot side-by-side with the video-chat meeting window, allowing the clinician to access the information in real-time.

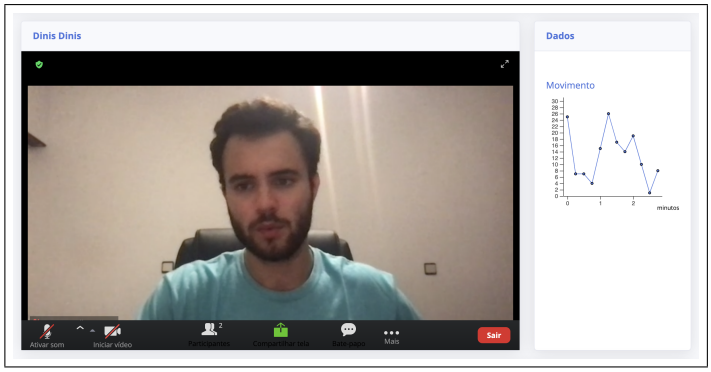


Figure 1: Teleconsultation page during a meeting with the head movement's chart.

1 Introduction

Depression is currently the major contributor for inability to be productive in society [1] and the coronavirus pandemic has brought an increase in isolation, leading to higher susceptibility of developing depressive symptomatology in the general population. The way depression is diagnosed and treated is still based on periodic clinical consultations, held between the patient and the doctor. The doctor's assessment of the patient's mental status is influenced by his clinical experience and his personal capacity. Naturally, this method is susceptible to human error from the doctor's side. This approach can benefit from biomarkers to help guide diagnosis, thus allowing doctors to more accurately detect and assess the severity of relevant symptoms.

Moreover, psychomotor retardation is common in people with depressive disorder [2, 3], thus being one of the nine symptoms that are part of the Diagnostic and Statistical Manual of Mental Disorders 5th edition (DSM-5) criteria for major depressive disorder (MDD) diagnosis [4]. It can be manifested as slowed speech, dysfunctional cognition and decreased movement. The movement is usually decreased in hands, legs, torso and head [5]. Typically, the amplitude and velocity of head movement are the metrics used to assess the difference between healthy and depressed patients. These have been showed to be decreased when depression severity was high [6].

Currently, there is a large government investment in digital health transformation. The Plano de Resolução e Resiliência (PRR) financial package plans a 300M€ investment in this sector. According to the President of Serviços Partilhados do Ministério da Saúde (SPMS) [7], one of the initiatives is the SNS24 desk, where it will be possible to carry out teleconsultations and provide support for telemonitoring prescribed by doctors. The goal is to have a desk in each Junta de Freguesia and thus provide access to every portuguese resident.

Taking all this into account, this work presents the development of a teleconsultation system that supports the diagnosis of depressive disorder with head movement's analysis.

This work was supported by Portuguese funds through FCT (Fundação para a Ciência e Tecnologia) through the projects reference UIDP/50009/2020 (Programático) and through the reference UID/EEA/50009/2019, LARSyS - FCT Plurianual funding 2020-2023.

2 Depression Assessment Systems

Recently, some attempts have been made to develop systems that provide objective measures to support clinical decision. For example, Zhou *et al.* (2015) [8] developed a system that assesses the mental health of the subject in real-time. Two experiments were performed in this work, one with healthy subjects, and the other with depressed patients. In the first experiment each subject had to read and respond to 6 sets of tweets. Each set had a predefined emotional type (positive, neutral, negative). While the task was being executed, the subject's video recording was being analysed, and one of the metrics the video analysed was head movement. The system was able to differentiate accurately when the patients were in positive or negative states. The second experiment obtained similar results to the first one. Nonetheless, this was done using a local recording, and not a teleconsultation platform that allows remote communication.

3 System Architecture

The proposed system is composed of three main components, namely: 1) a web app; 2) a database; and 3) the video processing.

3.1 Web App

The web app is the platform the doctor will use to have the teleconsultation with the patient. It was implemented using HTML, CSS and JavaScript. The SB Admin 2 Bootstrap template was used as an initial structure for the development.

Typically, the doctor will start the meeting, obtain an invitation and then request the patient to enter the teleconsultation. Similarly to a Zoom meeting, the teleconsultation will continue for an unlimited time frame until the doctor decides to end it. The teleconsultation is possible because the Zoom Meetings SDK was integrated in the app. For this, Zoom requires creating a developer account in Zoom Marketplace and then register it as a JSON Web Tokens App (JWT). The JWT is the method that Zoom API uses to establish a secure authentication when a request to start a meeting is made. It is generated from the API key and API secret, which are unique and given when the app is registered. Finally, the teleconsultation app during the meeting with the head movement results can be seen in Figure 1.

3.2 Database

The database is used to store the results from the head movement's analysis in each teleconsultation. The service used to create the database is Cloud Firestore, a NoSQL cloud database for mobile, web, and server development, from Firebase and Google Cloud.

The data is stored in a folder structure, and every user has a document where all the consultations analysis results are stored. This means that the values obtained during every teleconsultation are available for consultation by the clinician after the meeting ends.

3.3 Video Processing

The video processing component explains the structure that was created to acquire the teleconsultation video and then analyse it. The analysis is done locally as there is a Python service running in the background of the computer that is being used by the doctor for the teleconsultation. This service has to be started manually on the computer's terminal. Furthermore, every doctor that wants to use the platform will need to have the appropriate environment to run the files in the background of their own computer, including the dlib, MSS and OpenCV python libraries.

In the main loop, the process goes from frame capture to biomarker analysis sequentially and uninterruptedly. The frame is obtained by selecting a portion of the teleconsultation video screen, and then captured using the MSS library. Every time a frame is captured, the head movement's analysis is done. Moreover, as Python allows multithreading, the results of the biomarkers analysis will be sent every 15 seconds to the database in parallel using a new thread, and thus the main loop is not affected. The workflow can be observed in Figure 2.

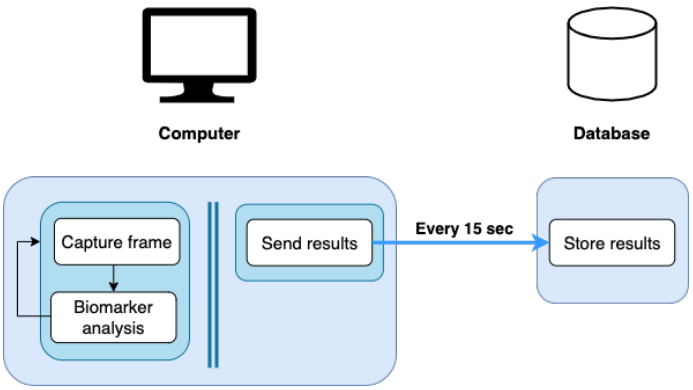


Figure 2: Video processing workflow.

4 Head Movement Analysis

The head movement analysis is the part where we take a series of frames and extract a metric that may be clinically relevant. The chosen metric is the average velocity every 15 seconds.

In order to measure this, a process is followed. First, the face of the subject is predicted using the dlib frontal face detector. Then, 64 landmark points are detected from the square identifying the face, using dlib landmark detector. These points can be indexed to identify specific areas of our face such as eyes, nose, mouth, eyebrows and these areas can then be used to extract biomarkers. The landmark point 28 in the nose was chosen as the head's center of mass. This process can be seen in Figure 3. Subsequently, the velocity of head's movement can be calculated by differentiating the positions of two consecutive frames of the nose landmark point, and the units are pixels/frame. As we need scalar values to be able to provide relevant medical information, the L2 norm of the vector velocity was computed. In order to obtain the values that will be visualized, the velocities obtained for every frame in the 15 seconds interval are averaged and sent to the database. These results are then shown to the clinician in a line chart, as shown in Figure 1.

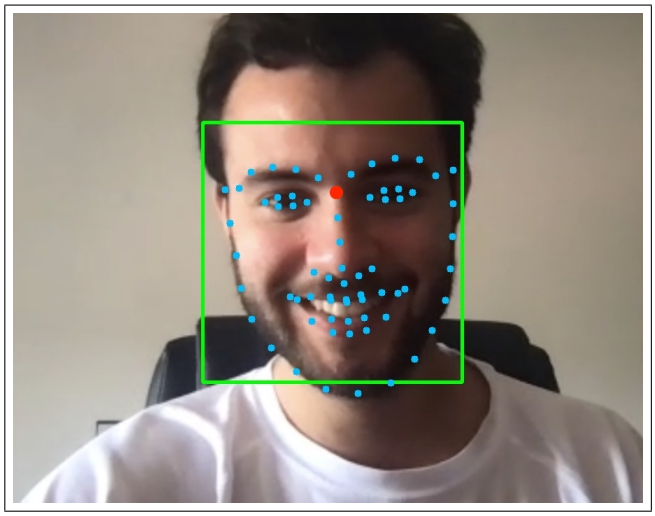


Figure 3: Overview of nose landmark point (red dot) detection. Face detection (green square). Remaining landmark points (blue dots).

5 Conclusions and Future Work

The main problem this work tackles is the fact that the psychiatric approach on depressive disorder is still mostly based on an interaction between patient and doctor that is periodic, very spaced in time and that uses only subjective measures for diagnose and follow-up. Until now, a platform that integrates patient's objective information to support psychiatrists' decisions during teleconsultation had never been developed. As the platform main goals were accomplished, this work is a step forward in filling the gaps.

Even though the chosen measure is interesting, it would be better if a scale was built that provides the doctor with an immediate universal measure of the patient's behavior, instead of just showing the values for average velocity. For instance, this scale could have values from 0 to 5, where 0 means no movement and 5 means a lot of movement.

References

[1] World Health Organization. Depression and other common mental disorders: global health estimates. Technical documents, 2017.

[2] J. F. Greden and B. J. Carroll. Psychomotor function in affective disorders: an overview of new monitoring techniques. *Am J Psychiatry*, 138(11):1441–1448, Nov 1981.

[3] D. J. Widlöcher. Psychomotor retardation: clinical, theoretical, and psychometric aspects. *Psychiatr Clin North Am*, 6(1):27–40, Mar 1983.

[4] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-V)*, chapter Humor Disturbances. American Psychiatric Publishing, 2013.

[5] J. S. Buyukdura, S. M. McClintock, and P. E. Croarkin. Psychomotor retardation in depression: biological underpinnings, measurement, and treatment. *Prog Neuropsychopharmacol Biol Psychiatry*, 35(2): 395–409, Mar 2011.

[6] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald. Nonverbal Social Withdrawal in Depression: Evidence from manual and automatic analysis. *Image Vis Comput*, 32(10):641–647, Oct 2014.

[7] Tsf. Em alta voz - luís goes pinheiro - presidente do spms - serviços partilhados do ministério da saúde, Jun 2021. URL `https://www.tsf.pt/programa/em-alta-voz/emissao/luis-goes-pinheiro---\presidente-do-spms---servicos-partilhados-do-ministerio-da-saude-13868752.html`.

[8] Dawei Zhou, Jiebo Luo, V. Silenzio, Yun Zhou, Jile Hu, G. Currier, and Henry A. Kautz. Tackling mental health by integrating unobtrusive multimodal sensing. In *AAAI*, 2015.

Face Detection and Alignment Using On-the-Wild Multispectral Images

Pedro Roque Martins
pedro.roque.martins@tecnico.ulisboa.pt
Military Academy, Lisbon, Portugal,
Instituto Superior Técnico, Universidade de Lisboa, Portugal

José Silvestre Silva
jose.silva@academiamilitar.pt
Military Academy & CINAMIL, Lisbon, Portugal,
LIBPhys-UC, Coimbra, Portugal

Alexandre Bernardino
alex@isr.tecnico.ulisboa.pt
Institute for Systems and Robotics (ISR), Lisbon, Portugal
Instituto Superior Técnico, Universidade de Lisboa, Portugal

Abstract

Face detection is a critical step for surveillance and access control applications. Coupled with face alignment, it supports face recognition/authentication applications in non-cooperative (on-the-wild) environments where facial poses are neither centred nor aligned at the time of image capture. The use of deep learning for image classification in computer vision has changed the paradigm in face detection and alignment, which has allowed a marked improvement in its accuracy. Because face analysis systems typically operate in the visible spectral band, it is in this spectral band that networks are trained. This paper describes some existing face detection and alignment methods. It tests their application in different spectral bands and in an on-the-wild environment to verify their generalization capability and select the best methods.

1 Introduction

Face detection is an essential step for tasks such as face recognition, face editing and face tracking. Although significant progress has been made during the last decade with the rise of deep learning, these methods still struggle with factors inherent to uncontrolled environments (e.g. variation of facial pose). Thus, that accurate and efficient face detection in natural conditions remains an open challenge [1].

Face detection, in conjunction with face alignment, aims to detect the faces presented in the input image and identify facial landmarks so that facial images are centred, aligned, and are equally sized. In the same way that one normalizes a set of feature vectors by centring them at zero or placing them on a unit scale before training a machine learning model, facial images can also be normalized, which facilitates further analysis. Since face detection algorithms detect faces in rectangles without rotation in the image, a face landmark detection algorithm is needed to apply a rotation so that the face is aligned on the horizontal axis, using, for example, the imaginary eye line. Thus, the procedure of face detection and alignment module consists in, given an image, identifying the different faces present, extracting the facial landmarks, and processing the image to produce facial images where the face is centred and aligned.

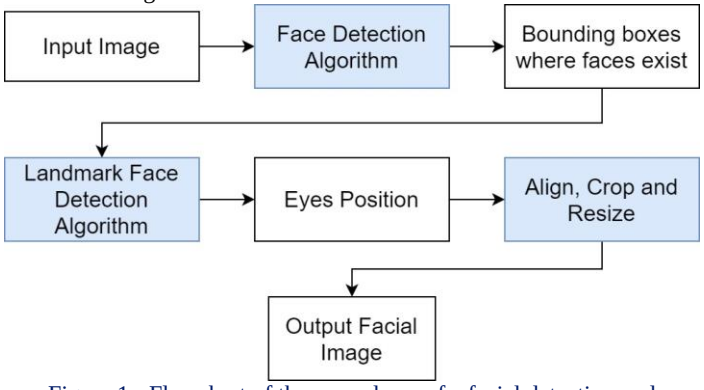


Figure 1 - Flowchart of the procedures of a facial detection and alignment module.

Like all tasks involving facial images, face detection and alignment have been studied mainly in the visible band, and the databases in this spectral band are where most machine learning-based face detection and face landmark detection algorithms are trained.

Multispectral images allow a facial analysis system to obtain facial features that would be impossible only with the spectral band of the visible. An example of this is the ability of the infrared spectral band to obtain images in low-light environments or even overcome occlusions such as smoke and fog. It is then interesting to analyze their generalization ability for the different spectral bands in an uncontrolled environment.

2 Methods

The face detection algorithms studied in this work are based on SSD (single-shot multibox detector), a deep learning architecture for object detection [2]. The basic idea of SSD is to define default bounding boxes using small convolutional filters on the feature map. At prediction time, the network generates scores for the presence of each object category in each predefined box and produces adjustments to the box to match the shape of the object. In this work, three SSD based methods are tested: (i) the S3FD algorithm [3], (ii) the facial detection deep neural network of OpenCV [4], and (iii) the DSFD algorithm [5]. The S3FD network has contributions to better cope with scaling variations with a single deep network. The DSFD network uses a feature enhancement module to extend the single-shot detector to dual-shot detector, obtaining more robust and discriminable features.

As for the facial landmark detection algorithms, the DLIB library's 68 landmark network, adapted from Khazemi and Sullivan [6], and Bulat's 2D-FAN [7], also with 68 landmarks were tested. The latter one uses an Hour-Glass [8] based architecture to estimate the human pose. Both networks receive an image of a person and produce, as output, the position of the different facial landmarks around the face.

All the algorithms were trained in databases that only contain images in the spectral band of the visible. To achieve data normalization, it is necessary to (i) rotate the image to align the eye line with the horizon, (2) crop the image to centre the face image, and (iii) resize the image so that all output images have the specified dimensions.

Qualitative and quantitative tests were performed. The qualitative tests allowed visualization of how the algorithms work in different spectral bands and poses, while the quantitative tests provided numerical values of the algorithms' performance.

3 Results and Discussion

3.1 Dataset

Qualitative evaluation methods use images obtained at the Military Academy to visualize the behaviour of face detection and facial landmarks detection algorithms. These images are in the Visible and Long Wavelength Infrared (LWIR) band, over 3 distinct poses.

For quantitative evaluation, a subset of the TUFTS [9] database was used, which has facial images in the visible band, Near Infrared (NIR) and LWIR of 113 people, with 9 different poses, with different lighting conditions. Since it was only possible to perform labelling for the bounding boxes and not for the face landmarks, only numerical results were obtained for the face detection algorithms.

Table 1 - Number of images per spectral band in the subset with pose variation of TUFTS [9].

Spectral Band	Number of images
Visible	3406
NIR	3444
LWIR	998

3.2 Face Detection

Regarding the qualitative results presented in Figure 2, all algorithms produced similar results in the visible band. This was expected since they were all trained in databases of the spectral band of the visible. In the LWIR spectral band, a failure of the OpenCV network was observed in the second facial pose, where it can't detect any face. In addition, when OpenCV and S3FD detect the faces, there is a variation in the rectangle area when compared to the visible spectral band. The

DSFD maintained the same results, being a good indicator of its ability to extract characteristics even in the LWIR spectral band.

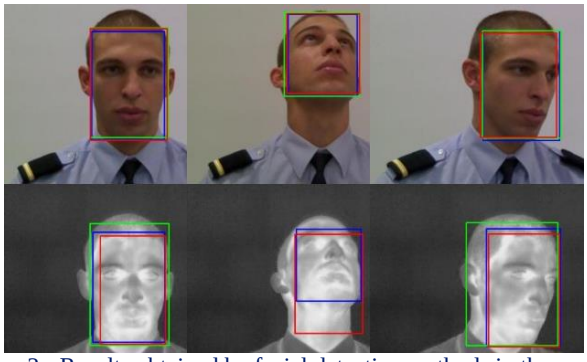


Figure 2 - Results obtained by facial detection methods in the spectral bands of Visible (above) and LWIR (below). S3FD-red, DSFD-blue, OpenCV-green.

The quantitative results are presented in Table 2. It can be observed that the OpenCV network results are lower than the others, especially in infrared bands. In the comparison of results between the S3FD network and the DSFD, we observe very similar results in the spectral band of the visible and NIR. However, the results in LWIR are about 8 percentage points better. We observe that the DSFD network maintains a very high accuracy for the different spectral bands, thus being the indicated network for face detection in a multispectral facial analysis system.

Table 2 – Accuracy of the different face detection algorithms in the TUFTS database.

	Accuracy at different spectral bands (%)		
Method	**Visible**	**NIR**	**LWIR**
OpenCV	99.18	90.36	77.66
S3FD	99.88	100.00	90.78
DSFD	99.88	99.97	98.79

3.3 Landmark Detection and Facial Alignment

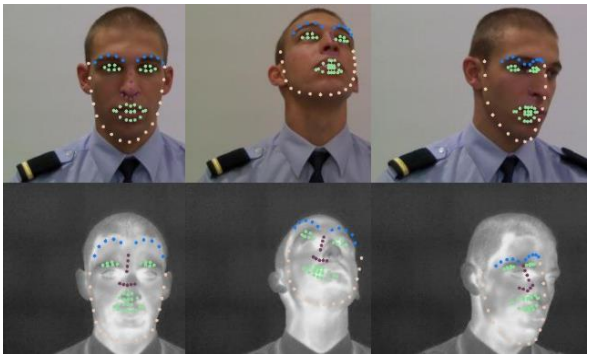


Figure 3 - Results obtained by Dlib in the spectral bands of Visible (above) and LWIR (below).

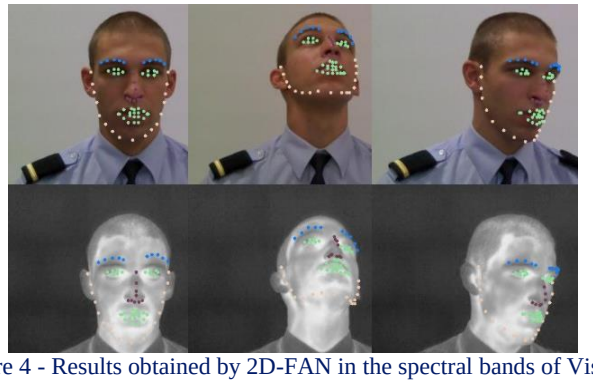


Figure 4 - Results obtained by 2D-FAN in the spectral bands of Visible (above) and LWIR (below).

The results for face landmark detection are shown in Figures 3 and 4. For the more challenging poses, we can see (Figure 3) that the DLIB network fails, even in the visible band, as it tends to maintain the shape of a near-frontal face. One possible cause of this behavior is the fact that the face landmark detection model was trained in a dataset without significant variations at the pose level. The DLIB network reveals even more difficulties in the spectral band of LWIR (see Figure 3).

As for 2D-FAN, it reveals a good extraction of landmarks in any of the poses, including in the LWIR band, where the results are pretty similar to those obtained in the visible band (Figure 4). This network, unlike DLIB, was trained on a database with pronounced pose variations (including profile images), which is one of the reasons why it gets better results.

After the face detection with DSFD and landmark face detection with 2D-FAN, the image processing phase took place, which aligned the imaginary eye line of all detected faces with the horizontal, centred the faces in the images and resized the cropped windows to the same size. The alignment effect is more noticeable on the rightmost facial image. This normalization of the facial images can help a multispectral facial recognition system in an uncontrolled environment (*on-the-wild*), where faces can be presented in several different poses.

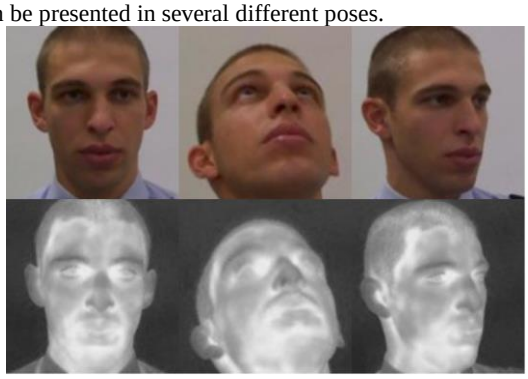


Figure 5 - Results obtained by the proposed facial detection and alignment module in the spectral bands of Visible (above) and LWIR (below).

4 Conclusions

The current analysis of several methods of facial detection and facial landmark detection revealed two main points. First, networks that are trained in diversified databases, with different poses and facial expressions, have better performance. Second, the methods that achieve good results in the visible spectrum also reveal a good performance in other spectral bands. The analysis of several methods enabled us to select the best face detection and alignment module, which is a crucial step in a multispectral facial analysis system on-the-wild.

Acknowledgements

This work was supported in part by the Military Academy Research Center (CINAMIL) under project Multi-Spectral Facial Recognition and by FCT with the LARSyS – FCT Project UIDB/50009/2020.

References

[1] S. Minaee et al., “Going Deeper Into Face Detection: A Survey,” CoRR, 2021.
[2] W. Liu et al., “SSD: Single Shot MultiBox Detector,” Lect. Notes Comput. Sci., 2016.
[3] S. Zhang et al., “S3FD: Single Shot Scale-Invariant Face Detector,” IEEE ICCV, 2017.
[4] G. Bradski, “The OpenCV library,” Dr Dobb’s J. Softw. Tools, 2000.
[5] J. Li et al., “DSFD: Dual shot face detector,” IEEE CVPR, 2019.
[6] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," IEEE CVPR, 2014.
[7] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," IEEE ICCV, 2017.
[8] A. Newell et al., “Stacked hourglass networks for human pose estimation,” Lect. Notes Comput. Sci., 2016.
[9] K. Panetta, et al., "A Comprehensive Database for Benchmarking Imaging Systems," IEEE PAMI, 2018.

An Initial Approach to Self-Supervised Underwater Fish Detection

Ricardo J.M. Veiga[1]
rjveiga@ualg.pt
Jorge Semião[2]
jsemiao@ualg.pt
João M.F. Rodrigues[1]
jrodrig@ualg.pt

[1] LARSyS & Instituto Superior de Engenharia
Universidade do Algarve
Faro, Portugal

[2] INESC-ID & Instituto Superior de Engenharia
Universidade do Algarve
Faro, Portugal

Abstract

Our world is an ever-changing entropic organism influenced by all the cogs and gears that compose it. With the oceans covering most of the surface of our planet, the way we interact with it and its abundant fauna influence our nature's balance. Conjoining the real-time information of the recognition of local fish species, with their abundance, gauge, and stress factors across different coastal locations, it is possible to develop a synergy with the local fisheries, which will prevent over-fishing, and also promote a sensible fishing response, while allowing the monitoring of invasive species. Our presented uses deep neural networks and self-supervised learning to detect fishes, it automatically adapts to different underwater conditions without human input, allowing to skip the main time-consumption task of any deep-learning detection project, and introducing a scalable solution to coastal monitoring. The prove of concept implemented presents very good results, 79.4% average precision on the dataset used.

1 Introduction

The necessity of monitoring our underwater *livestock* is not an original idea of our current age. Since the first time any human fed on any of the diverse *menu* the oceans had to offer, that the need to track the best places to fish became almost an obligation. With our progress and evolution, came new techniques that enhanced our ability to fish, and with technology, we finally reach the ability to drain our oceans of all of their resources [4]. The need for equilibrium arose, therefore, the need for control and monitoring of our oceans [11].

Coastal' surveys are usually done by fishery and oceanography science, by the use of trawls. This manual type of method to gather analytics on the fish abundance is extremely invasive to the local fauna's habitats and doesn't reflect real-time data information, alas, it is dependent on a team of specialists in terms of resources and time. However, with the advances of machine learning and big data analytics, it is now possible to automatically monitor the underwater marine fauna, removing the need for personal surveys, or manual examination of hours of footage by marine biologists.

Through the adaptation of popular deep learning pipelines for object detection, authors have already shown the potential of computer vision and artificial intelligence to tackle this challenge [6, 7, 8, 10]. Although, the presented approaches rely on analysis of data from the same locations, with the model developed being custom made for that specific local, which is not scalable.

In this paper, we propose a different approach to the data annotation process through the use of self-supervised learning. Through the use of transfer learning and pseudo-labelling, we removed the need for human labelling for training data preparation. Our current approach can easily readapt to a new location, and increase its accuracy over time while using tracking to reduce type I errors.

2 Underwater Fish Detection

Considering our goal of a scalable solution for the detection of marine fauna in real-time on multiple locations, environments, and conditions, either by the use of an edge device, cloud processing, or local processing, we chose the YOLOv4 (You Only Look Once) algorithm [2] for our application, due to its versatility across complex scenes while still maintaining high accuracy and efficiency.

Following the traditional transfer-learning object detection training algorithm, we used pre-trained weights for the convolutional layers. These weights were previously trained on the Common Objects in Context (COCO) dataset [5], and detect its 80 classes. Using an already-trained network gives us the advantage of faster training using fewer resources, while still maintaining similar performance. However, on a more abstract view, for our case, we are only detecting one type of class: fish, which a (more) simpler network trained from scratch would suffice, and using a multi-class network would normally be considered excessive. Nonetheless, this method allowed us to achieve a vaster generalization, with multiple distance detections, and lower type I and type II errors.

2.1 Data

The acquisition and annotation of data for training is a long and tedious process of most machine learning projects regarding non-standardized object classes from the main used datasets. The data collected should reflect the application, and maintain a uniform level of labelling quality.

For our goal, we obtain footage from our coastal waters from the Centro de Ciências do Mar (CCMAR). We have 21 unlabelled videos, with a sum of 695 minutes, and an average of 33 minutes each. Also, the recording of these videos starts onboard and finishes randomly. This footage is from different locations of our coastal area, with various underwater conditions. We also have a ground truth selection of 756 unseen images from different locations, as can be seen in Fig. 1 top row.

With our pre-trained weight prepared to detect the COCO classes, we needed to find an initially labelled data train to detect the class: *fish*. The main dataset used for training fish detectors and classifiers is the Fish4Knowledge dataset [3], although, due to its lower resolution, and with the increased power of remote cameras, we decided to use the OzFish dataset[1], which increased the performance of our approach due to higher quality images [9]. Fig. 1, two bottom rows shows a sample of the annotated part for object detection.

OzFish is a public dataset of the Australian fish species. After the exploratory data analysis, we remained with 1800 frames with approximately 45k bounding boxes. It is important to note that these annotations were generated through an outsourcing cloud processing platform, therefore, we evaluated deeper the data and found erroneous bounding box annotations. Although this makes the data unreliable, using our approach, we were able to still achieve a high accuracy rate.

3 Method

To prepare our footage for detection, we had to define which part of it is underwater. We used traditional computer vision methods to divide the videos. Through the histogram information, combined with movement

Figure 1: Top row, ground truth annotations of CCMAR videos. Bottom two rows, sample of OzFish dataset object detection images

analysis, we isolated the underwater footage from the overwater, although, we still had the transition while the camera and baiter is being lowered to the bottom of the sea. Using background subtraction analysis through adaptive Gaussian mixtures, and histogram anomaly analysis, we were able to define the moment the structure hits the bottom floor. The inverse process was used to also isolate the retrieval of the structure from the footage.

Type I & II errors are false-positives and false-negatives detections, respectively. To minimize these errors, we use negative images retrieved from the CCMAR's footage, as can be seen in Fig. 2. The addition of these unlabelled images to the training dataset, in a similar amount to the same, helps to cancel the unwanted classes from the pre-trained weights. Lowering the type I & II errors, without deteriorating the network, allowing for a later pruning. With the data prepared from the OzFish dataset, we add the negative images and began the transfer-learning method with a high resolution for the input images of the network at 960 pixels for width and height. This higher resolution allows for higher precision during the self-supervised learning process.

From the OzFish trained model, we perform data mining on the CCMAR's videos to generate a pseudo-dataset through the use of unsupervised pseudo-labelling. Here, we initiate a loop of self-supervised learning, where the annotations are automatically analysed across continuous frames through intersection over union (IoU) tracking, and the resulting images are used to continue training the model, with a progressive blur applied to the non-detected areas of the images. To improve generalization in different scenarios, the *random* selected frames from the videos to have conditions and rules to prevent repeated, or overly similar, data.

While we are training more and more data that was labelled by itself on one class: fish, we are training two very different classes, we are also training the background. For object detection, the background information is extremely valuable. This is the main reason to blur the remaining image, or we would be *teaching* our network that the background also included the false-negative fishes. As soon as the accuracy reaches the desired threshold, in our case we defined >90%, we perform inference on the videos and the annotated images became an autonomous dataset and are used to train the different models according to the application - edge device, local processing, cloud processing.

4 Results

The metrics used to evaluate the results are divided in two. One for the ground truth of the already seen footage used for training, and another for the never seen before images annotated by a marine biologist. For the first metric, we achieve an average precision of 92.03%, with an average IoU of 69.08%. On the second metric, using images from unseen locations, we obtained an average precision of 79.38%, with an average IoU of 69.56%.

Figure 3 presents some of the achieved results. Even in dense scenarios or obscured with floating sand and debris, we were able to detect the majority of the present fishes. When we added tracking, we were able to maintain a more stable detection across the footage.

5 Conclusions and Future Work

We were able to obtain great results on similar locations to the ones that we trained, and also good results on different unseen environments. With the present approach, we were able to remove the most time-consuming phase of the traditional custom object detection process while still retaining strong results.

As future work, we are devising a way to improve the validation of the automated annotations through temporal attention models. We are also developing a platform for easier ground truth label validation from the marine biologist using our method as a backend.

Figure 2: Sample of negatives images used during training.

Figure 3: Sample of detections from our current model (see text).

Acknowledgements

This work was supported by LARSyS - FCT Project UIDB/50009/2020, and project KTTSEADRONES, project 0622_KTTSEADRONES_5_E, financed by POCTEP - Interreg VA Spain-Portugal. We would also like to thank CCMAR - Centro de Ciências do Mar, for their support, availability, and access to their data and knowledge.

References

[1] University of Western Australia (UWA) Australian Institute of Marine Science (AIMS) and Curtin University. Ozfish dataset - machine learning dataset for baited remote underwater video stations, 2019. URL `https://apps.aims.gov.au/metadata/view/38c829d4-6b6d-44a1-9476-f9b0955ce0b8`. Accessed Dec. 08, 2020.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[3] Robert B Fisher, Yun-Heh Chen-Burger, Daniela Giordano, Lynda Hardman, Fang-Pang Lin, et al. *Fish4Knowledge: collecting and analyzing massive coral reef fish video data*, volume 104. Springer, 2016.

[4] Belinda Gallardo, Miguel Clavero, Marta I Sánchez, and Montserrat Vilà. Global ecological impacts of invasive species in aquatic ecosystems. *Global change biology*, 22(1):151–163, 2016.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[6] Shasha Liu, Xiaoyu Li, Mingshan Gao, Yu Cai, Rui Nian, Peiliang Li, Tianhong Yan, and Amaury Lendasse. Embedded online fish detection and tracking system via YOLOv3 and parallel correlation filter. In *OCEANS 2018*, pages 1–6. IEEE, 2018.

[7] Delphine Mallet and Dominique Pelletier. Underwater video techniques for observing coastal marine biodiversity: a review of sixty years of publications. *Fisheries Research*, 154:44–62, 2014.

[8] Md Moniruzzaman, Syed Mohammed Shamsul Islam, Mohammed Bennamoun, and Paul Lavery. Deep learning on underwater marine object detection: A survey. In *Int. Conference on Advanced Concepts for Intelligent Vision Systems*, pages 150–160. Springer, 2017.

[9] Dominique Pelletier, Kévin Leleu, Gérard Mou-Tham, Nicolas Guillemot, and Pascale Chabanet. Comparison of visual census and high definition video transects for monitoring coral reef fish assemblages. *Fisheries Research*, 107(1-3):84–93, 2011.

[10] Herman Stavelin, Adil Rasheed, Omer San, and Arne Johan Hestnes. Marine life through you only look once's perspective. *arXiv preprint arXiv:2003.00836*, 2020.

[11] Xinting Yang, Song Zhang, Jintao Liu, Qinfeng Gao, Shuanglin Dong, and Chao Zhou. Deep learning for smart fish farming: applications, opportunities and challenges. *Reviews in Aquaculture*, 13 (1):66–90, 2021.

Sketch-to-Photo Matching Enforcing Realistic Rendering Generation

Leonardo Capozzi[1,2]
leonardo.g.capozzi@inesctec.pt
João Ribeiro Pinto[1,2]
joao.t.pinto@inesctec.pt
Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt
Ana Rebelo[2]
arebelo@inesctec.pt

[1] Faculdade de Engenharia
Universidade do Porto
Porto, Portugal
[2] INESC TEC
Porto, Portugal

Abstract

The use of forensic sketches to locate suspects is a challenging task. These sketches are posted on public spaces, social media and the news with the hope that someone recognizes the suspect. Recent methods present some limitations, as they do not use end-to-end networks or/and do not offer other alternatives in case the matching process fails, such as providing a photo-realistic representation of the sketch. This paper presents a method that combines a conditional generative adversarial network (cGAN) and a pre-trained face recognition network, optimised as an end-to-end model. This method is able to retrieve a list of potential suspects, and it simultaneously provides an intermediate realistic representation of the sketch. Evaluation on the CUFS and CUFSF databases shows that the proposed method outperforms state-of-the-art methodologies in most tasks, and that forcing a photo-realistic rendering of the sketch only results in a slight performance decrease.

1 Introduction

Over the years, the use of deep learning has brought a lot of advancements in pattern recognition and computer vision tasks, such as face recognition. Recent methodologies report significantly higher accuracy in the matching process when using deep learning [8, 10]. However, the use of real photos in face recognition is much easier than the use of a forensic sketch, since a forensic sketch might not be a very accurate representation of the suspect, as it was drawn using descriptions from eye witnesses [9].

Recent state-of-the-art methods have used convolutional neural networks (CNN) to perform sketch-to-face matching [2, 3, 5, 7], however many of these do not use end-to-end approaches, which can limit the performance of the model by causing dissonance between separately optimised blocks.

Other methodologies include the generation of a photo-realistic representation of the sketch, which can be useful for manual identification in case the matching process fails. They use adversarial approaches based on CycleGANs [5] and cGANs [2, 7].

This work tackles two important aspects that have not been addressed in the literature. The first aspect is the use of an end-to-end model which is jointly optimised, avoiding the performance limitations associated with the use of separate processes. The second aspect is enforcing the end-to-end model to generate an intermediate photo-realistic representation of the input sketch that can be used by law enforcement in manual matching processes.

The proposed methodology is composed of a cGAN and a matching CNN that are optimised in an end-to-end fashion. When trained, the model receives a sketch and returns a feature vector that can be used for matching using simple distance metrics. The model also generates an intermediate latent representation which is realistic and similar to the corresponding real photographs.

2 Proposed Methodology

The proposed methodology is composed of two main parts, which are jointly optimised: a sketch-to-render generator and a matching network (see Fig. 1). The sketch-to-render generator receives a sketch and outputs a photo-realistic representation of the sketch that is similar to the real face of the person. The matching network receives the intermediate photo-realistic representation and outputs a feature vector that can be used for the matching process.

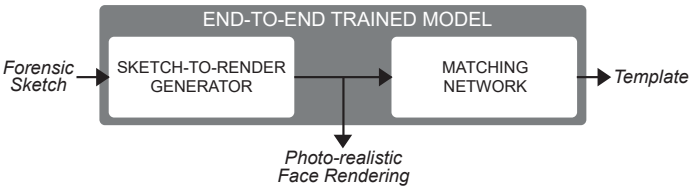


Figure 1: Overview of the proposed methodology (Taken from [1]).

2.1 Network Architecture

The generator consists of a U-Net, which is an encoder-decoder with skip connections that enable the transmission of information between corresponding levels of the network. It receives a sketch and outputs a photo-realistic rendering of the sketch [1].

The discriminator used in this work is a CNN adapted from the cGAN of pix2pix [4]. It receives as input the photo-realistic rendering of the sketch outputted by the generator and the corresponding sketch, and outputs a prediction on whether the photo-realistic rendering is a real image or generated image [1].

The matching network used in this work is a VGG-16 with pre-trained weights from VGG-Face [8]. It receives as input the photo-realistic rendering of the sketch outputted by the generator and outputs a feature vector. The sketch is matched to real photos from a database by computing the cosine distance between the respective feature vectors.

2.2 Loss

The loss function used for training is composed of several losses. The first component comes from the loss of the cGAN, and it is responsible for generating photo-realistic images. It can be written as the following:

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x}[\log(1 - D(x,G(x)))], \quad (1)$$

where x is a sketch and y is the corresponding ground-truth photo. The generator (G) tries to minimize the loss and the discriminator (D) tries to maximize it.

The generator should also try to mimic the real photograph of the person in the sketch, therefore we add a second loss term:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y}[||y - G(x)||_1], \quad (2)$$

which minimizes the $L1$ norm of the difference between the real image (y) and the generated image ($G(x)$).

The identity of the generated realistic rendering needs to match the identity of the ground-truth image. Hence, we add a third component to the loss function:

$$\mathcal{L}_{match}(G) = \mathbb{E}_{x,y}[||V(y) - V(G(x))||_2]. \quad (3)$$

The weights of the VGG-Face network (V) are frozen, since we are using the pretrained weights. The matching loss only adjusts the weights of the generator.

Combining all the loss components, the final loss function becomes:

$$\mathcal{L}(G,D) = \min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda_1 \mathcal{L}_{L1}(G) + \lambda_2 \mathcal{L}_{match}(G). \quad (4)$$

Table 1: Matching accuracy on the CUFSF and CUFS datasets, using different methods to enhance the sketch (r.r.g.: realistic rendering generation) (Taken from [1]).

Method	CUFSF			CUFS		
	R-1	*R-5*	*R-10*	*R-1*	*R-5*	*R-10*
Sketch	49	77	88	47	80	90
pix2pix	53	83	92	52	79	86
IPMFSPS [6]	**74**	94	97	**80**	**95**	**97**
HFFS2PS [2]	-	-	-	36	69	-
Proposed (with r.r.g)	54	87	96	44	77	86
Proposed (without r.r.g)	**74**	**98**	**99**	59	85	91

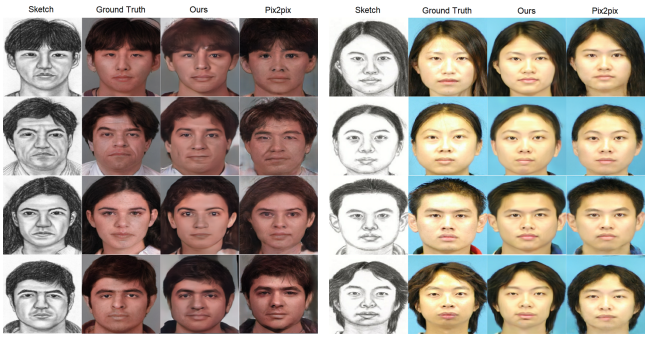


Figure 2: Images generated by our method using the CUFSF dataset (on the left); Images generated by our method using the CUFS dataset (on the right). (Taken from [1]).

3 Experimental Settings

The proposed methodology was trained using sketch-photo pairs from the CUHK Face Sketch database (CUFS) [11] and the CUHK Face Sketch FERET dataset (CUFSF) [11, 12].

The photos of CUFSF from the FERET database were colorized using the DeOldify API (Available on: https://github.com/jantic/DeOldify). This was done to uniformize the datasets and to allow the model to generate color images.

The sketches and photos were transformed so that the position of the eyes was consistent in each sketch-photo pair, in order to improve the quality of the generated images and the accuracy of the matching process.

The model was trained on two experimental settings. The first one used the previously mentioned loss function, in order to generate intermediate realistic renderings and to have a high matching accuracy. The second setting measured the impact that removing the loss terms $\mathcal{L}_{cGAN}$ and $\mathcal{L}_{L1}$, which promote the generation of a realistic image, had on the matching accuracy. For more details refer to [1].

4 Results and Discussion

To measure the performance of the matching process we computed the rank-N accuracy (R-N) on the test sets of CUFS and CUFSF. In this case, we consider $N \in \{1, 5, 10\}$.

The results in Table 1 show that the proposed method is superior or aligned to the alternative methods on all considered ranks. However, when the photo-realistic rendering is dismissed, the matching accuracy increases significantly, showing a trade-off between matching performance and the realism of the rendering, which could be tuned for specific scenarios.

Examples of photo-realistic renderings, the corresponding sketches, ground-truth photos and images generated using the pix2pix method can be seen in Fig. 2. Visually, we can confirm that the generated renderings look realistic, and similar to the ground truth photos. Considering the performance of the model, maintaining the photo-realism is a positive aspect of the proposed methodology.

5 Conclusion

This paper proposes an end-to-end-method for sketch-to-photo matching that enforces a photo-realistic rendering of the input sketch. Upon evaluation, the matching process showed that the proposed method is superior or aligned to state-of-the-art methodologies, and the generation of intermediate face renderings offered realistic results. The matching results improved when disregarding realistic face renderings, showing a trade-off between matching accuracy and the realism of the generated image. Further efforts should be devoted to improve the realistic rendering generation, in order to allow for a more diverse range of face characteristics, such as hair, eyes, and skin color. We believe these efforts would improve the results of both the realistic renderings and the matching process.

Acknowledgements

This work was partially funded by the Project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership, and within the PhD grants "SFRH/BD/137720/2018" and "2021.06945.BD". Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

References

[1] Leonardo Capozzi, Jaime S. Cardoso, Ana Rebelo, and Joao Pinto. End-to-end deep sketch-to-photo matching enforcing realistic photo generation. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (CIARP'21)*, 2021.

[2] W. Chao, L. Chang, X. Wang, J. Cheng, X. Deng, and F. Duan. High-fidelity face sketch-to-photo synthesis using generative adversarial network. In *ICIP*, pages 4699–4703, 2019.

[3] S. M. Iranmanesh, H. Kazemi, S. Soleymani, A. Dabouei, and N. M. Nasrabadi. Deep sketch-photo face recognition assisted by facial attributes. In *IEEE BTAS*, pages 1–10, 2018.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[5] H. Kazemi, M. Iranmanesh, A. Dabouei, S. Soleymani, and N. M. Nasrabadi. Facial attributes guided deep sketch-to-photo synthesis. In *WACVW*, pages 1–8, 2018.

[6] Y. Lin, S. Ling, K. Fu, and P. Cheng. An identity-preserved model for face sketch-photo synthesis. *IEEE Signal Processing Letters*, 27: 1095–1099, 2020.

[7] Uche Osahor, Hadi Kazemi, Ali Dabouei, and Nasser Nasrabadi. Quality guided sketch-to-photo image synthesis. *arXiv*, 2020. 2005.02133.

[8] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[9] Sourav Pramanik and Dr. Debotosh Bhattacharjee. An approach: Modality reduction and face-sketch recognition. *arXiv*, 2013.

[10] Mei Wang and Weihong Deng. Deep face recognition: A survey. *arXiv*, 2018.

[11] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 31. IEEE, 2009.

[12] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, 2011.

From Captions to Explanations: Towards In-Model Unsupervised Natural Language Explanations

Isabel Rio-Torto[1,2]
isabel.riotorto@inesctec.pt
Luís F. Teixeira[1,2]
luisft@fe.up.pt
Jaime Cardoso[1,2]
jaime.cardoso@inesctec.pt

[1] INESC TEC
Porto
Portugal

[2] Faculty of Engineering of the University of Porto
Porto
Portugal

Abstract

The growing importance of the Explainable AI field has resulted in the proposal of several methods for producing visual heatmaps of the classification decisions of deep learning models. However, visual explanations are not enough since different end-users have different backgrounds and preferences. Natural language explanations are inherently understandable by humans and, thus, can complement visual explanations.

Therefore, we introduce a novel transformer-based architecture to tackle the generation of natural language explanations without their direct supervision. Preliminary experiments show the potential of the approach and shed light on how it can be improved.

1 Introduction

The same way a visual explanation differs from a segmentation map [5], so does a textual explanation differ from a caption [1]. While the caption constitutes a detailed description of an image, ideally mentioning all the existing objects, an explanation should only refer to (part of) objects that are relevant for a certain classification outcome.

In the literature, several works [1, 2, 3] approach the problem of generating natural language explanations for classification problems as traditional supervised image captioning tasks, where the model learns to produce some ground-truth explanations. For example, in Park et al. [3] the authors collect datasets by asking humans to provide natural language explanations for each image. However, we argue that for an explanation to reflect the decision process made by a model, the model cannot be directly trained to generate said explanation. By doing so, we would simply obtain a captioning model, when in fact what we want is to be able to produce textual explanations for the predictions of a classification model. Therefore, our main focus lies in generating such explanations without direct supervision, i.e. without ever learning from ground-truth explanations.

2 Proposed Approach

Given the shortage of datasets where a clear distinction between the captioning and the classification tasks is present, we started by creating a toy dataset that allows us to perform faster prototyping and easily evaluate the proposed solutions without needing specific expert knowledge. The dataset is composed of images containing triangles and/or squares of different colours. An image is labelled as positive if there is at least one triangle on the left of the leftmost square, and is considered negative otherwise. As such, there is a clear contrast between a caption, which ideally should account for all the polygons and their relative positions, and an explanation for the classification label, which should only need to identify the presence of the leftmost square and focus on the objects on its left. Therefore, each image in the dataset is also accompanied by 3 possible captions and an explanation. We generated 2000 images for training, 600 for validation and 200 for testing; each image has at least 2 and at most 4 polygons.

According to [1], an explanation should be simultaneously image and class relevant. Keeping this in mind, we propose the architecture depicted in Fig. 1. Its main building blocks are an encoder-decoder transformer using a Vision Transformer (ViT) as encoder and an auto-regressive language model as decoder. The hidden states from both modules are fed to linear projection layers and their outputs are concatenated and given to a Multilayer Perceptron (MLP) with 3 Linear+ReLU layers for the final classification. To obtain the ViT hidden states we use its CLS token, since it constitutes a latent representation of the whole image that can be used for classification. For the decoder using the CLS token is only possible if it follows a BERT-like architecture. Furthermore, Zhang et al. [6] concluded that using max pooling achieved better results when compared to using the CLS representation, something that we also concluded in preliminary experiments. Thus, we also employed a max pooling strategy over all the token representations of the last layer of the decoder.

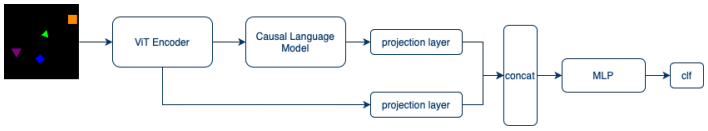


Figure 1: Diagram representing the proposed transformer-based architecture for self-explanatory classification with natural language explanations.

Table 1: Image captioning results obtained on the toy dataset.

Decoder	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDER	SPICE
GPT-2	0.591	0.550	0.495	0.434	0.422	0.566	0.034	0.517
DistilGPT-2	0.600	0.562	0.509	0.449	0.430	0.572	0.041	0.525
BERT (uncased)	0.901	0.821	0.733	0.648	0.434	0.707	**2.589**	0.556
BERT (cased)	**0.903**	**0.833**	**0.755**	**0.677**	**0.446**	**0.717**	2.477	**0.573**

The main rationale behind this architecture is to guarantee both image and class dependence by, respectively, generating text from the image and using that text to influence the classification outcome. Similarly to what was proposed in previous work developed in the scope of the in-model generation of visual explanations [4], if the explanations directly contribute to the classification, then the classification loss will alter the explanations accordingly, making them reflect what is important for the classification task. Notwithstanding, the classification loss is not a strong enough supervisory signal to ensure that coherent text is produced. Thus, we introduce a first training step in which the transformer is trained for supervised image captioning and then transfer the resulting weights to the second training step, where the whole network is fine-tuned for classification.

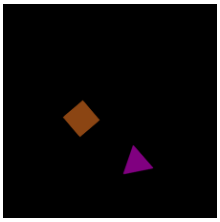
Finally, to regularise the training we randomly ignore one of the input modalities, i.e. sometimes both image and text embeddings are used, sometimes only the text counterpart is used and sometimes only the image representation is used. When a single modality is used we repeat that modality projection layer's resulting vector to be able to keep the structure of the concatenation block.

3 Results and Discussion

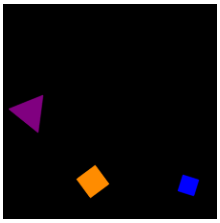
3.1 Image Captioning

For every experiment we used an ImageNet pre trained ViT encoder with 224x224 resolution. Several decoders were explored, including GPT-2, DistilGPT-2 and the base versions of BERT with and without case sensitivity. All models were trained with cross entropy for 20 epochs with a batch size of 8 and an initial learning rate of 5e-5, linearly decayed. During training we consider sequences until 95 tokens and during inference sentences can have 100 tokens at maximum. Finally, we use beam search decoding with 4 beams.

From Table 1 we can conclude that using BERT as a decoder yields better results across all metrics. This can be explained by the fact that GPT-2 is originally trained for open-end text generation, so it becomes more difficult to learn when to stop producing tokens. In fact, we verified that with GPT-2-based models the EOS token is never reached and the produced captions always present the maximum allowed number of tokens, while with BERT the produced sentences have different lengths; after a certain number of tokens BERT stops the generation and GPT-2 starts producing incoherent text until its output is truncated. In the end, these unnecessary extra sentences hinder GPT-2-based models' performance. Naturally, the cased version of BERT performs slightly better than the uncased version simply because our ground-truth captions are them-



Label: negative
Predicted class: negative
Generated caption: There are a total of 2 polygons, of which one is a square and of which 1 is a triangle. The square is brown. The triangle is purple. The brown square is on the right of the purple triangle.
GT explanation: The leftmost figure is the brown square.
Generated explanation: There are a total of two shapes, all are triangles. The triangles are orange and white. The orange triangle is on the right of the white triangle.



Label: positive
Predicted class: positive
Generated caption: There are a total of three polygons, of which 1 is a triangle and of which 2 are squares. One square is orange and the other is blue. The triangle is purple. The rightmost figure is the blue square and the leftmost figure on the right is the purple triangle. The orange square is in the middle.
GT explanation: The purple triangle is on the left of the orange square.
Generated explanation: There are a total of 2 polygons, of which one is a triangle and of which 1 is a square. The triangle is brown. The square is blue. The brown triangle is on the left of the square.

Figure 2: Natural language explanation results obtained. For each example we present the ground-truth (GT) and predicted classification labels, the generated caption, as well as the GT and generated explanations. The generated explanations refer to training with all three alternatives (image-only, text-only and multimodal) with equal probability. In red are highlighted mismatches between the generated text and the corresponding image and in green are parts of the text that correctly match the image.

Table 2: Objective evaluation of the generated natural language explanations. The "text", "image" and "multimodal" (concatenation of image and text embeddings) columns refer to the different modalities used during training and their respective probabilities. The model which was trained using all modalities with equal probability performs slightly better.

text	image	multimodal	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDER	SPICE
1.0	0	0	0.113	0.077	0.047	0.025	0.177	0.200	0.000	0.206
0	1.0	0	0.147	0.105	0.077	0.059	0.208	0.241	0.000	0.271
0	0	1.0	0.143	0.100	0.065	0.041	0.193	0.244	**0.002**	0.192
0.5	0	0.5	0.143	0.105	0.068	0.043	0.202	0.236	0.000	0.226
0	0.5	0.5	0.130	0.087	0.055	0.031	0.200	0.199	0.000	0.213
0.5	0.5	0	0.129	0.079	0.046	0.022	0.177	0.209	0.000	0.193
0.33	0.33	0.33	**0.163**	**0.124**	**0.095**	**0.075**	**0.228**	**0.272**	0.000	**0.319**

selves cased. As such, we opt for the model trained with the cased version of BERT as our initialisation strategy for the second training phase.

3.2 Natural Language Explanations

For the second training phase, models were trained for 10 epochs with a batch size of 16 and an initial learning rate of 1e-4, linearly decayed. We experimented with randomly changing which embeddings were given to the MLP, either only the image, or only the text or concatenating both (which we call Multimodal). We also experimented with combinations of the previous alternatives, for example switching between multimodal and text-only with 50% probability each or using all three with 33% probability each. In terms of classification accuracy all models perform well within a small number of epochs, achieving 99% accuracy.

Regarding the performance of the system in terms of the generated explanations, there is a slight improvement of the model trained with all three modalities with equal probability. Nevertheless, the results of all models can be greatly improved (see Table 2 and Fig. 2). The explanations differ from the captions obtained in the first training stage, which is expected considering that in this second phase they are directly influenced by the classification. However, the explanations lose their image relevance, for example, the generated text describes more polygons than the ones actually present or switches their colour. In the majority of images of the positive class the explanations only mention the existence of two polygons, which is consistent with the original rationale that to identify the positive class one only needs to pay attention to the leftmost square and to one triangle on its left. Furthermore, in every image described as having two polygons some variation of the sentence "There are a total of 2 shapes, of which one is a triangle and of which 1 is a square." occurs. There is also a recurrence of the sentences "The triangle is brown." and "The brown triangle is on the left of the square." in about 20% of the positive instances. This suggests that the network might be learning that is it sufficient to say that one triangle is on the left of the square in order for the image to be classified as positive.

4 Conclusion and Future Work

We proposed a novel transformer-based architecture to produce natural language explanations for classification decisions without direct supervision of those explanations. However, preliminary experiments show that using only the classification loss might not be enough for this task. In fact, the generated explanations lack image relevance. So, as future work, we will explore the following hypothesis: if a caption describes an image then it should be possible, to some degree, to reconstruct an image from its caption. Analogously, if an explanation describes which parts of the image are relevant for a classification outcome, then it should be possible, to some extent, to reconstruct those important parts from said explanation, i.e. it should be possible to generate a visual explanation from a textual explanation. Thus, we will include a reconstruction transformer responsible for generating an image from the corresponding text, which will allow us to explore some sort of cyclic consistency similarly to CycleGAN: the first encoder-decoder transformer would generate text from the input image, while the second would generate an image from text.

5 Acknowledgements

This work was partially funded by the Project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership and by the PhD grant "2020.07034.BD".

References

[1] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European conference on computer vision*, pages 3–19. Springer, 2016.

[2] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding Visual Explanations. In *Lecture Notes in Computer Science*, volume 11206 LNCS, pages 269–286. 2018.

[3] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8779–8788. IEEE, jun 2018.

[4] Isabel Rio-Torto, Kelwin Fernandes, and Luís F Teixeira. Understanding the decisions of CNNs: An in-model approach. *Pattern Recognition Letters*, 133(C):373–380, may 2020.

[5] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017.

[6] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

Evaluation of the Accuracy of Pose Estimation Based on Relative Pose

Francisco Lourenço
francisco.rlourenco@gmail.com
Helder Araújo
helder@isr.uc.pt

Institute of Systems and Robotics
Department of Electrical and Computer Engineering
University of Coimbra

Abstract

In this paper we describe an approach for the evaluation of the estimation of absolute pose. The proposed approach is based on the estimation of the relative pose and on the computation of a metric based on the relative rotation and relative translation between objects. When multiple objects are present in a scene and if only the camera moves, their relative poses remain constant between frames. Using the absolute poses of the objects, the relative poses in each frame can be estimated and their variation can be used to evaluate the algorithms. One of the advantages of use of the relative pose is that it can be applied even if a ground truth pose is not available, e.g., in pose estimation approaches without object recognition.

1 Introduction

Absolute pose of both RGB and RGB-D images has been addressed using machine learning approaches. A comprehensive review of 6DoF object pose estimation can be found in [1]. 6DoF pose estimation can be performed at two different levels:

- Instance-level - When a method is said to work at the instance-level, it means that, to estimate the 6DoF pose of an object, such a method will estimate the pose of a known object instance, i.e., an object used in the training phase.

- Category-level - As opposed to instance-level, category-level approaches deal with unseen objects. Instead of precisely identifying the object they want to look for, these methods work with categories, e.g., cars, bicycles, boxes, toys. For example, while at the instance-level, the methods know precisely the brand, model, and color of a car whose pose they want to estimate, at category-level, the only information the methods have is that they should look for a car. Estimating 6DoF object pose at the category-level is usually a more complex problem to solve, but methods that work at this level generalize better than at the instance-level, as it might be possible to estimate the pose of a broader range of unseen objects instances.

Furhtermore 6DoF pose estimators can be divided into two main categories:

- 3D bounding box detectors - These methods work at the category-level and do not estimate the 6DoF pose directly but, instead, they fit a 3D bounding box to the object. To do so, these methods produce oriented 3D BBs (Bounding Boxes) centered at some point $x = (x, y, z)$, size $d = (d_w, d_h, d_l)$ with orientation (θ_y). The bounding box can then be extended to the 6DoF space where the pose can be recovered.

- Full 6DoF pose estimators - Directly estimate the 3D translation and 3D rotation. Typically, these methods work at the instance-level. However, recent proposed full 6DoF pose estimators such as [2] address the category-level problem.

Several approaches for pose estimation exist. [3] is an instance-level/ classification/ full pose estimator. This method was originally designed to estimate poses from RGB images but the authors also present in their work a variant where they adapt it by opening a depth channel and consequently estimating the 6DoF pose from RGB-D inputs. [4] is an instance-level/template matching/full pose estimator. This method uses a SVM and templates that are used for object detection and, if a 6DoF pose is assigned to the templates, these can vote for the 6DoF pose of an object instance. [5] is an instance-level/template matching/full pose estimator. This approach is quite simple and focuses on solving the problem of template matching-based methods where the templates are built online from the RGB-D outputs and require physical interaction of a human operator or a robot with the environment.

In the work described in this paper we decided to evaluate the approach DenseFusion [6], which is an instance-level/ regression/ full pose estimator. The core idea of this approach is to embed and fuse RGB values and point clouds at the per-pixel level. The goal is to fully leverage RGB and depth information. The use of the depth information as an extra channel of the RGB image is not considered since these data are defined in different domains. In DenseFusion a heterogeneous architecture is proposed, where RGB data and depth information are processed individually and then densely fused at the per-pixel level, where each extracted feature will vote for a 6DoF pose.

2 Methodology

This study aims to infer the 6DoF pose estimator accuracy in terms of the relative pose of the objects present in the image. Having multiple images of the same scene but acquired from different positions of the camera, the absolute poses of the objects change. However, their relative poses with respect to each other remain constant. Hence, by measuring how these relative poses change between frames, it is possible to infer the quality of a pose estimation model, not only by measuring how these relative poses change between frames, but also from the ground truth poses given on a dataset.

2.1 Proposed Metric for 6DoF Pose Estimator Precision Measurements

Having a set of frames of size N, each containing k objects, with $N > 1$ and $k > 1$, and the respective poses, and if we consider that the camera acquires images of all objects in every frame, an object can be randomly chosen as a reference, which will be denoted as object o_r. Then, the poses of the other objects relative to the reference object are calculated. This is done for all frames and will result in $k-1$ relative transformations per frame. These transformations are equal in all frames under the assumption that the objects did not move during image acquisition. Therefore, the differences between corresponding relative poses in different frames can be estimated to characterize the accuracy of the approach.

If we write a 3D transformation as:

$$[R|t] = \begin{bmatrix} r_1 & r_2 & r_3 & t_x \\ r_4 & r_5 & r_6 & t_y \\ r_7 & r_8 & r_9 & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

We can compute the relative transformation between the reference object o_r and object j in frame n as follows:

$$n[R|t]_{o_r j} = n[R|t]_{o_r}^{-1} \cdot n[R|t]_j \tag{2}$$

This is done for all frames and, when all relative 3D transformation matrices $n[R|t]_{o_r j}$ are available, the error between these matrices is measured between relative transformations in the frames by calculating the product of the inverse of matrix $(n-1)[R|t]_{o_r j}$ by matrix $n[R|t]_{o_r j}$. Essentially, we are calculating the relative transformation between relative transformations and, if these are the same, the result should be an identity matrix of size $4*4$. Hence, the error will be the Frobenius norm of the difference between the resulting matrix and an identity matrix, as expressed in the following equation:

$$n\varepsilon_{[R|t]_{or\,j}} = ||I_4 - (n-1)[R|t]_{or\,j}^{-1} \cdot n[R|t]_{or\,j}||_f \quad (3)$$

Where $n\varepsilon_{[R|t]_{or\,j}}$ is the j^{th} element of error vector $n\varepsilon_{[R|t]}$ (vector containing all $k-1$ transformation errors of consequent frames) between frames n and $n-1$. This will yield N-1 error vectors of size $k-1$ that, in the end, are averaged into a single precision score. This score will be denoted as $\tau_{[R|t]}$ and can be seen as how much the poses "jitter" between frames. The smaller it is, the better. The use of the relative coordinate transformation instead of the Frobenius norm of the difference between matrices is due to the fact that this measure corresponds to the pose "difference" measured on the manifold. The object to be used as a reference can be any of the objects from the training set. It is selected "a priori" allowing for the automatic computation of the metric.

2.2 The Proposed Metric as a Loss Function

This metric can also be used as a loss function. In fact the similarity between relative poses can also be used for training a machine learning model. Therefore, if we consider equation 5, and substitute the relative poses by an estimated and a target pose, we can infer their similarity. This function can then be back-propagated and hence used to update the weights of a neural network.

For that purpose the proposed metric can now be rewritten as:

$$L = ||I_4 - [R|t]_{target}^{-1} \cdot [R|t]_{estimated}||_f \quad (4)$$

For DenseFusion, the equation can be written as follows, where a confidence score is also added so that the model can learn in a self-supervised fashion:

$$L = \frac{1}{N_{Features}} \sum_i (c_i \times ||I_4 - i[R|t]_{target}^{-1} \cdot i[R|t]_{estimated}||_f - \omega \times log(c_i)) \quad (5)$$

When compared with the ADD(-S) based loss function, a significant advantage of this metric is that it does not require the objects 3D models. For that reason, the proposed loss function can be less computationally intensive, thus improving training time.

2.3 Results

In this work the YCB-Video dataset was used to evaluate the accuracy. This dataset was proposed in[7] and it provides accurate 6D poses of 21 objects observed in 92 videos with 133,827 frames.The poses estimated by DenseFusion and the ground truth poses of the YCB-Video dataset were used to evaluate the accuracy. However, this metric can be used for any dataset or 6DoF pose estimator, as long as the images acquired correspond to the same scene with multiple object instances that did not move while the images are obtained.

===	YCB-Video	DenseFusion Original
$\tau_{[R\|t]}$	0.0011	0.0887
====		

Table 1: Relative pose study results (accuracy).

Figure 1: YCB ground-truth (1st row); DenseFusion (2nd row) poses.

The results shown here were obtained for the evaluation videos from the YCB-Video dataset, totaling 20738 frames.

As expected, the YCB-Video Dataset labels are the most precise poses. These results can be seen as how much jitter there is on the estimated poses between consequent frames. Visually, if we fit the 3D models to the RGB images, a low score will result in a smoother representation, while with high scores, the models look like they vibrate during the video. So, this score can be seen as how much "vibration" it is on the projected 3D models. Figure 1 illustrates this.

References

[1] Caner Sahin, Guillermo Garcia-Hernando, Juil Sock, and Tae-Kyun Kim. A Review on Object Pose Recovery: from 3D Bounding Box Detectors to Full 6D Pose Estimators. jan 2020.

[2] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-PACK: Category-level 6D Pose Tracker with Anchor-Based Keypoints. oct 2019.

[3] Eric Brachmann, Frank Michel, Alexander Krull, Michael Yang, Stefan Gumhold, and Carsten Rother. Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image. In *CVPR*, volume 2016-Decem, pages 3364–3372. IEEE, jun 2016.

[4] Reyes Rios-Cabrera and Tinne Tuytelaars. Discriminatively Trained Templates for 3D Object Detection: A Real Time Scalable Approach. In *ICCV*, pages 2048–2055. IEEE, dec 2013.

[5] S. Hinterstoisser and V. Lepetit and S. Ilic and S. Holzer and G. Bradski, K. Konolige, and N. Navab. *Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes*, volume 7584 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[6] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martin-Martin, Cewu Lu, Li Fei-Fei, and Silvio Savarese. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. In *CVPR*, volume 2019-June, pages 3338–3347. IEEE, jun 2019.

[7] You Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. Nov 2017.

An Exploratory Study on ECG Biometric Bias Using Compression Algorithms

João Carvalho
joao.carvalho@ua.pt
Susana Brás
susana.bras@ua.pt
Armando J. Pinho
ap@ua.pt

IEETA
University of Aveiro
Aveiro, Portugal Portugal

Abstract

On most machine learning applications an open issue is how to properly deal with unbalanced datasets and avoid bias towards some classes in favor of others. This is also a problem on biometric identification systems, where such a behavior from the system would allow some individuals to deceive it, being identified as someone else. In order to solve this problem, it is important to understand why some individuals are more prone to force this behavior to the system. We use an open database and perform different experiments on one-to-one subject biometric identification, using different pairs of participants. The setup procedure is to maintain all training data in one of the subjects and gradually trimming the data for the other one to see how robust is the classification process. Our exploratory study shows that some participants can be successfully identified even when there is only 10% of the training data available – an highly unbalanced problem. Given that the classification algorithm used does not correct the imbalance of the data, this is highly surprising and deserves to be further explored, as it suggests that some participants "biometric signature" might be available in just a small portion of data.

1 Introduction

There is a big interest in biometric systems nowadays, for diverse purposes. One of the signals that seems to be of great interest for biometry is the electrocardiographic (ECG) signal, as it conveys desirable characteristics for biometric identification (universality, uniqueness, measurability, acceptability and circumvention avoidance), which are not the case on some alternatives based on image recognition, as they might be easily counterfeited.

On previous studies, we have began researching how compression-based methods can be used for this purpose, using a measure of relative similarity called the Normalized Relative Compression (NRC) [5]. However, biometric identification systems, like any other artificial intelligent systems, make mistakes. Those mistakes, when examined closely, do not seem to be random by nature, as they tend to be made when trying to identify a specific subset of participants. In this study, we start an exploratory study on what errors are made by such a system so that, in the future, they may be prevented.

We will use a mechanism to force data imbalance between the classes, in order to evaluate how it affects the performance of the biometric identification system. We will use the previous results as a baseline to explore what mistakes are made by the system, as it seems that the biometric signature of individuals differs in nature. We will then discuss how this might affect real systems and what can be done to make such systems more robust.

2 Methods

In order to preprocess the signals for compression, we used a Butterworth low-pass filter of 5th order at 30Hz cut-off frequency.

Given that the ECG signal suffers from baseline wander, it makes more sense to operate on its derivatives (see [5] for more details). Therefore, after the preprocessing step, a Lloyd-Max quantizer [7] with an *alphabet size* of 17 was used on the consecutive derivatives of the signal.

In order to compute the number of bits required to perform the relative compression, we have used the same type of model as the one used in [5]: an extended-alphabet finite-context model[1] (xaFCM) – a more general version of an FCM [4] with a context size $k = 35, \alpha =$'auto' and $d = 1$ (for more information on the last two parameters see [4]). It is important to mention that we do not want to achieve real compression – we just want to compute measures of relative similarity (the NRC). More details for the classification scheme used can be found in [5].

3 Experimental Results

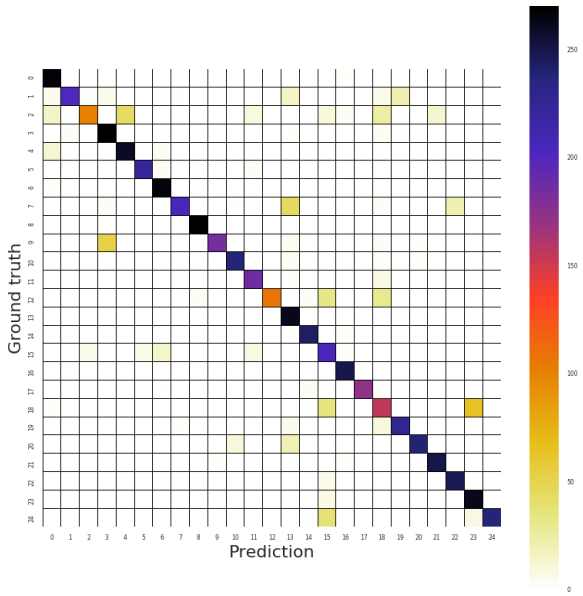


Figure 1: The confusion matrix obtained for the biometric identification when using the whole database with the parameters described (alphabet size of 17, $k = 35$, $alpha =$'auto' and $d = 1$).

To perform this study, we used an ECG open database that has been previously used for biometric identification [3, 4, 5] and is publicly available for research purposes[2]. The ECG was collected from 25 participants on three different sessions and was sampled at 1000 Hz. On each session, participants were exposed to a certain stimuli (fear, happiness and neutral). For more details, see [2].

On previous studies, where we used the same database, the test made was to use two out of the three sessions of data per participant as the training data (reference data) and the other session as the test data (target data). Each test segment contained exactly 10 seconds (see [3] for more information). On the previous study, we could achieve an accuracy of approximately 89.3% [5]. The confusion matrix for that problem can be seen in Fig. 1. This was the ground floor for the tests that we have done on the current work.

However, since the purpose of this study was to explore the impact of the data imbalance available for training each class (subject) in the biometry results obtained and how it might provide bias towards some participants, we opted for a simpler setup: instead of using all the subjects on each test, we used two subjects per test and performed a one-to-one classification – *i.e.* we performed the biometric identification between pairs of individuals at each time. This way it was easier to notice how decreasing the amount of data available for the reference of participants, forcing the problem to be unbalanced, impacts the biometric identification results. Given the time required for each experiment and also the exploratory nature of the experiment, we have decided to do only 50 of these tests.

[1] The source code was implemented using Python 3.6 and is publicly available under the GPL v3 license at `https://github.com/joaomrcarvalho/xafcm`

[2] `http://sweet.ua.pt/ap/data/signals/Biometric_Emotion_Recognition.zip`

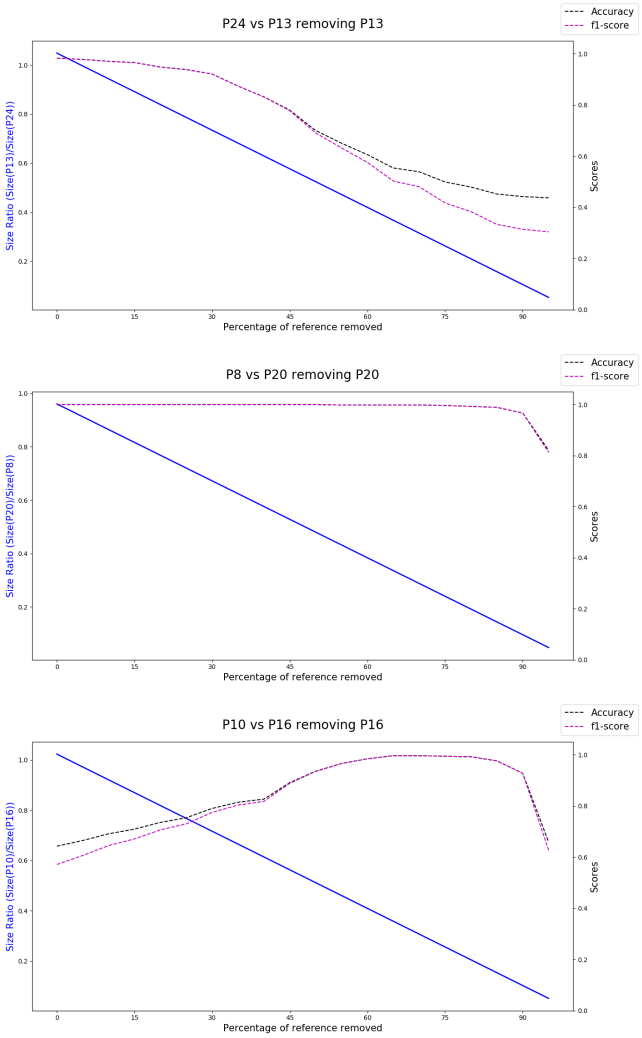


Figure 2: **(Dashed-lines)** Measuring the accuracy and f1-score as the percentage of the reference of participant one of the participants is trimmed from the end. On the top plot, both performance measures decrease slowly as expected; on the middle plot it is possible to see an example where even with only 10% of the reference data the subjects can be easily distinguished; on the bottom plot is an example where there is another kind of weird behavior by the system – instead of decreasing the performance measures as on of the reference gets trimmed, the opposite occurs (until a certain point). **(Blue line)** Measuring the ratio of reference size available for one participant and the other – the further away from 1, the higher the data imbalance.

It is important to notice that we never used data from the same recording/session for training and testing. Since the ECG signal changes substantially from one session to another, using data from the same session for the training and test would lead to over-optimistic results [3]. The reason for this is that there is an intra-variability for each participant from one day to another, which makes the biometric identification more challenging [1].

The basis for all tests is to pick two participants randomly, use all the available reference data for both participants and keep removing parts of the reference for the other participant to see how the performance measures for the biometric identification are affected as the data imbalance increases between the classes. Instead of measuring the amount of data we have for each class, we only care about the ratio between the data available for one class and the other – the further away from one, the higher the data imbalance. It is also worth mentioning that we measure the amount of data by the number of samples available and not the file size itself. We performed three different tests:

1. The beginning of the reference (training) data was removed, 5% at each time;

2. The end of the reference (training) data was removed, 5% at each time;

3. The training data was divided into 20 equally sized intervals and the first 5% of the data of each interval were removed.

The idea behind having three different tests was to evaluate if the results obtained were similar when removing data from the reference on different regions. Surprisingly, the results were similar on the three tests for almost all pair of participants evaluated. For that reason, the experiments suggest that the biometric results depend on the amount of data available for training each class and it does not seem to be related to a specific part of the signal.

Between all test combinations, three different categories of results were obtained, *i.e.*, three different types of behaviour were found on the experiments. In Figure 2 we show one result for each of those categories, using test number one (removing the beginning of the reference data) and reporting both the accuracy and f1-scores obtained as the reference data of one of the participants is removed. From those results we can see clearly the three different categories/behaviors:

(a) A decrease on the performance measures as the information available for the reference (training) decreases – what was theoretically expected. As the data available for the reference decreases, the problem becomes more unbalanced and tends towards the class for which more data is available. Example: top test from Figure 2;

(b) An almost perfect biometric identification between the two subjects, even when the problem starts becoming highly unbalanced – it is possible to notice that even without 80%-90% of the original reference data for one of the participants, the biometric identification is still successful most of the time. Example: middle test from Figure 2;

(c) Instead of the results getting worse as the problem becomes more unbalanced, they improve. This does not happen on most of the tests, but it is worth inspecting these specific cases to get some insights on what is the mechanism that causes this unexpected behavior – that might be useful for understand the bias on biometric identification. It is important to mention that this behavior is the opposite of what was expected. Example: bottom test from Figure 2.

Acknowledgments

This work was partially supported by national funds, European Regional Development Fund, FSE through COMPETE2020, through FCT, in the scope of the framework contract foreseen in the numbers 4, 5 and 6 of the article 23, of the Decree-Law 57/2016, of August 29, changed by Law 57/2017, of July 19; and in the scope of the projects UIDB/00127/2020 (IEETA/UA). João M. Carvalho acknowledges the doctoral grant from FCT, ref. SFRH/BD/136815/2018.

References

[1] F. Agrafioti, D. Hatzinakos, and A. K. Anderson. Ecg pattern analysis for emotion detection. *IEEE Transactions on Affective Computing*, 3(1):102–115, jan 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.28.

[2] Susana Brás and Armando J Pinho. Ecg biometric identification: A compression based approach. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5838–5841, aug 2015. doi: 10.1109/EMBC.2015.7319719.

[3] João M Carvalho, Susana Brás, Jacqueline Ferreira, Sandra C. Soares, and Armando J Pinho. Impact of the acquisition time on ecg compression-based biometric identification systems. In *Pattern Recognition and Image Analysis. IbPRIA 2017. Lecture Notes in Computer Science, vol 10255. Springer, Cham*, pages 169–176. Springer, Cham, jun 2017. doi: 10.1007/978-3-319-58838-4_19.

[4] João M. Carvalho, Susana Brás, Diogo Pratas, Jacqueline Ferreira, Sandra C. Soares, and Armando J Pinho. Extended-alphabet finite-context models. *Pattern Recognition Letters*, 112:49–55, sep 2018. ISSN 0167-8655. doi: 10.1016/J.PATREC.2018.05.026.

[5] João M. Carvalho, Susana Brás, and Armando J Pinho. Compression-based classification of ecg using first-order derivatives. In Paulo Cortez, Luís Magalhães, Pedro Branco, Carlos Filipe Portela, and Telmo Adão, editors, *Intelligent Technologies for Interactive Entertainment*, pages 27–36, Cham, 2019. Springer International Publishing. ISBN 978-3-030-16447-8.

[6] Nima Karimian, Paul A. Wortman, and Fatemeh Tehranipoor. Evolving authentication design considerations for the internet of biometric things (iobt). In *Proceedings of the Eleventh IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis - CODES '16*, 2016. doi: 10.1145/2968456.2973748.

[7] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, mar 1982. doi: 10.1109/tit.1982.1056489.

Evaluating the Performance of Zero-Shot Learning Methods using Low-Power Devices

Cristiano Patrício
cristiano.patricio@ubi.pt
João Neves
jcneves@di.ubi.pt

Departamento de Informática
Universidade da Beira Interior
NOVA LINCS
6201-001, Covilhã, Portugal

Abstract

Zero-shot recognition is more prone to be used in real-world scenarios when compared to traditional object recognition. Nevertheless, no work has evaluated the feasibility of deploying zero-shot learning approaches in these scenarios, particularly when using low-power devices. In this paper, we provide the first benchmark on the inference time of zero-shot learning, comprising an evaluation of state-of-the-art approaches regarding their speed/accuracy trade-off when performing in low-power devices. The results reveal that the visual feature extraction is the major bottleneck in the processing chain of the ZSL paradigm, but we show that lightweight networks can dramatically reduce the overall inference time without reducing the accuracy obtained by the *de facto* ResNet101 architecture. To foster the research and deployment of ZSL systems capable of operating in real-world scenarios, we release the evaluation framework used in this benchmark (`https://github.com/CristianoPatricio/zsl-methods`).

1 Introduction

Zero-shot learning (ZSL) emerged as a more realistic alternative to traditional object recognition, where the goal is to build computational models capable of identifying objects solely with a semantic description of the class, and without samples in the training phase.

In the last years, researchers have successfully exploited the advances on machine learning to boost the performance of this learning paradigm. At the same time, the effort put on creating standard evaluation protocols and benchmarks fostered the advances on the problem of ZSL. Thus, is not surprising that soon ZSL methods will be integrated into industrial solutions or end-user applications.

Most contributions on ZSL disregard the visual feature extraction phase by using precomputed features from standard Convolutional Neural Networks (CNN) architectures. The reduced size of these features significantly decreases the processing time of the feature classification in the ZSL processing chain, even when the classification algorithm has a high temporal complexity. For this reason, most works have focused solely on reporting accuracy, and few works have analyzed inference time of the proposed models. In [8], the authors perform an analysis on the complexity of the learning strategy. Ji *et al*. [2] report the processing time during both the training and inference phase, and provide the complexity of the proposed algorithm. Pan *et al*. [5] evaluate the inference time of the proposed method both on CPU and GPU, as well as the time required by competing approaches.

Nevertheless, no work has specifically evaluated the inference time of the overall processing chain of ZSL, neither the impact of using different architectures for the visual feature extraction phase. Also, the evaluation on low-power devices has not been considered yet. To the best of our knowledge, our work is the first benchmark on ZSL inference time in low-power devices, providing a comparative analysis of how different CNN architectures impact the speed/accuracy trade-off of these approaches.

2 Evaluation Methodology

We have selected six state-of-the-art ZSL methods, including ESZSL [6], SAE [3], DEM [11], f-CLSWGAN [10], TF-VAEGAN [4], and CE-GSZL [1]. The selected approaches cover the two major strategies in ZSL: (1) projection-based methods (ESZSL, SAE, and DEM), and (2) generative methods (f-CLSWGAN, TF-VAEGAN, and CE-GZSL).

This work was supported by the research grant UIDB/04516/2020/TRA/BIL/08 from the research unit NOVA Laboratory for Computer Science and Informatics (NOVA LINCS).

The evaluation of the six state-of-the-art ZSL methods considered in this study is carried out on the two most popular ZSL datasets, namely, *Animals with Attributes 2* (AWA2) [9] and *Caltech-UCSD-Birds* (CUB-200-2011) [7].

Experiments were performed in a desktop computer and two small low-power devices, namely a Raspberry Pi 4 Model B (R-PI 4B) and a Jetson Nano Developer Kit.

The multi-way classification accuracy (MCA) was adopted to assess the average per-class accuracy in the restricted setting [9]. In the generalized setting, the average per-class classification accuracy is determined on training (Y^s) and test (Y^u) classes, and the harmonic mean is obtained by $H = \frac{2 * acc_{Y^s} * acc_{Y^u}}{acc_{Y^s} + acc_{Y^u}}$, where acc_{Y^s} and acc_{Y^u} denotes the accuracy of seen and unseen classes, respectively.

3 Results and Discussion

3.1 ZSL Methods: Inference Time

The results of the inference time for classifying a single test image, considering 2,048 dimensional features, are reported in Table 1.

Method	Desktop	R-PI 4B	Jetson Nano
SAE	0.13±0.00	1.66±0.07	1.88±0.05
ESZSL	0.04±0.00	1.40±0.15	0.97±0.07
DEM	3.11±0.14	26.74±0.80	24.28±1.47
f-CLSWGAN	0.96±0.10	5.72±0.12	1.79±0.05
TF-VAEGAN	3.08±0.06	44.95±2.53	10.23±0.76
CE-GZSL	0.95±0.06	5.33±0.13	1.71±0.09

Table 1: Processing time (in milliseconds) of feature classification on CPU, R-PI 4B and Jetson Nano.

In general, ZSL methods are capable of classifying visual features of an image in less than 1ms when using CPU, and in less than 25ms when using low-power devices. Regarding the comparison between projection methods and generative methods, it is not possible to conclude which strategy is faster. Instead, the difference lies in the type of models used, as the methods based on deep learning models have a higher processing time.

3.2 Visual Feature Extraction: Inference Time

To analyze the time consumed by the visual feature extraction phase, which is usually performed using CNNs, we evaluate several CNN architectures with respect to the time consumed for obtaining the features of the top-layer pooling units, and using different hardware devices. The results are reported in Table 2.

Architecture	Desktop	R-PI 4B	Jetson Nano	Feat. Dim.	Size (MB)
MobileNet	**25.57**±3.17	310.52±9.80	**29.58**±3.14	1024	16
MobileNetV2	27.59±3.38	**297.63**±8.54	33.04±20.11	1280	14
InceptionV3	33.80±2.81	609.23±3.54	143.28±29.19	2048	92
NASNetMobile	39.67±2.35	370.10±5.79	111.81±20.40	1056	23
Xception	43.43±3.29	1081.18±11.07	155.75±23.21	2048	88
ResNet101	57.46±2.99	1639.46±76.71	-	2048	171
VGG16	59.73±4.10	2046.51±15.03	221.75±16.96	512	528
EfficientNetB7	86.54±0.85	2436.62±106.28	-	2560	256
NASNetLarge	95.67±2.98	2115.31±16.84	-	4032	343

Table 2: Processing time of the visual feature extraction phase on CPU, R-PI 4B and Jetson Nano.

The results show that visual feature extraction is significantly slower when compared with the inference time of ZSL methods (refer to Table 1), being the bottleneck in overall inference stage of ZSL. Also, the results evidence that the consumed time varies largely over different architectures, and only lightweight models are capable of providing an acceptable running time in a low-power computational devices such as Jetson

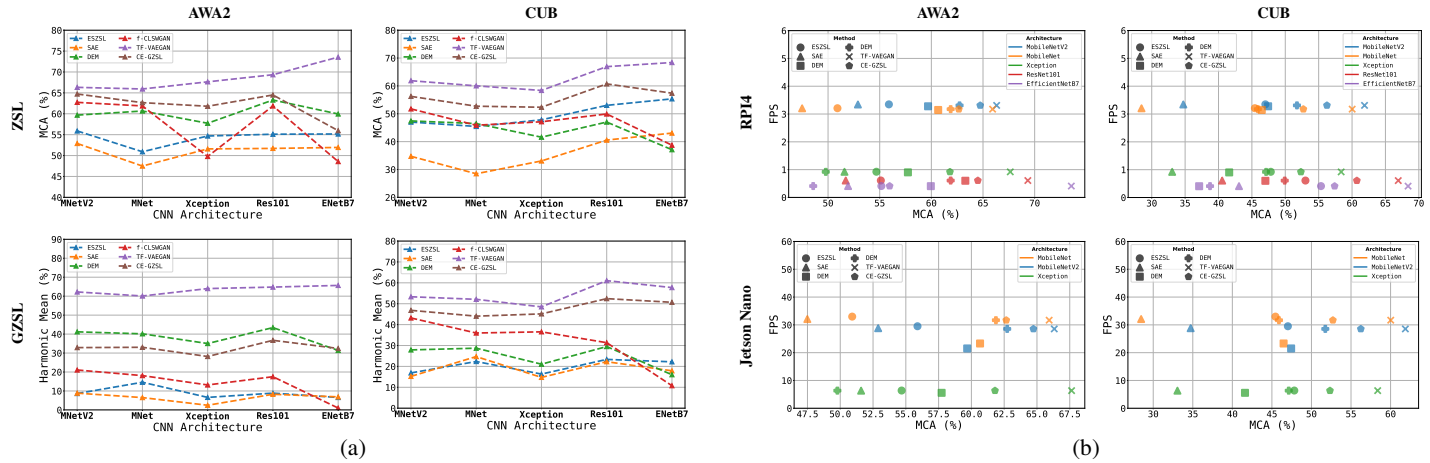


Figure 1: (a) Performance of ZSL methods with respect to the CNN architecture used. (b) Accuracy/Speed trade-off of ZSL methods with respect to the CNN architecture used.

Nano. However, these lightweight architectures are hardly used in ZSL applications, since the vast majority of the works adopt the ResNet101 architecture as the *de facto* model for benchmarking the accuracy of ZSL methods.

3.3 ZSL Accuracy: Impact of CNN Complexity

In order to understand the impact of using other CNN architectures in the accuracy, the ZSL methods considered in this study are re-trained using the features extracted from networks with varying complexity. To ensure a fair evaluation, the hyper-parameters of each approach are adjusted during training using a validation set. Finally, each method is evaluated in the datasets considered in this benchmark (AwA2 and CUB) under the restricted and generalized (GZSL) settings. The results are depicted in Figure 1a, and evidence that the features generated by lightweight architectures allow ZSL approaches to attain competitive results when compared with the *de facto* ResNet101 architecture. Also, it can be observed that, in general, lines are nearly horizontal, meaning that the performance of ZSL methods does not consistently improve with the complexity of the architecture used for feature extraction. These results suggest that the inference phase of ZSL methods can be speed up without compromising the accuracy of ZSL.

3.4 ZSL Speed/Accuracy Trade-off

We evaluate the performance of ZSL methods according to the MCA using two datasets commonly used in the field of ZSL (AwA2 and CUB) under the restricted setting. The overall inference time of ZSL methods is measured according to the frames per second (FPS) processed when using low-power computational devices. Also, to ensure a fair and comprehensive evaluation of the performance of these devices, the complexity of CNN models is reduced by using different network optimization techniques, such as layer fusion, matrix normalization, and the reduction of floating point precision (FP16/INT8). These optimizations are carried out using the NVIDIA TensorRT, allowing the creation of models compatible with the integrated GPU of Jetson Nano. Figure 1b depicts the results obtained, organized by dataset (columns) and the device used for inference (rows).

As expected, lightweight networks significantly decrease the processing time of ZSL inference phase. However, the throughput of these networks does not exceed 4 FPS when using Raspberry Pi 4B, decreasing its applicability in real-world scenarios. In contrast, Jetson Nano is capable of delivering 30 FPS using lightweight networks, which can be explained by the optimizations performed using TensorRT and the the use of a integrated GPU.

Regarding the comparison between lightweight architectures and the ResNet101, the top-1 accuracy decreases 4% in average considering the evaluated datasets, while the number of FPS increases from 0.6 to 3.3 in Raspberry Pi 4B, and 6.2 to 30.6 in Jetson Nano. This suggests that lightweight networks offer a compelling trade-off between inference time and accuracy, when compared to the *de facto* architecure used for benchmarking ZSL accuracy.

4 Conclusions

This work introduces the first benchmark on zero-shot learning regarding the processing time in the inference stage. The obtained results confirm that visual feature extraction is the major bottleneck in the pipeline of ZSL approaches, justifying thus the need for studying the impact of using different CNN architectures in the processing time of this phase. Accordingly, we measure the accuracy and processing speed of different ZSL methods when using architectures of varying complexity. Also, the evaluation was carried out using different hardware devices, to assess the feasibility of deploying ZSL methods in low-power computational devices. The results suggest that the use of lightweight architectures does not significantly decreases the accuracy of ZSL methods, while reducing dramatically inference time. Moreover, the analysis to processing time of low-power devices shows that the use of specialized hardware, such as low-power devices with integrated GPUs, can significantly reduce the processing time of visual feature extraction, enabling to perform the inference at 30 FPS.

References

[1] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. *arXiv preprint arXiv:2103.16173*, 2021.

[2] Zhong Ji, Yunlong Yu, Yanwei Pang, Jichang Guo, and Zhongfei Zhang. Manifold regularized cross-modal embedding for zero-shot learning. *Information Sciences*, 378:48–58, 2017.

[3] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3174–3183, 2017.

[4] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[5] Chongyu Pan, Jian Huang, Jianguo Hao, and Jianxing Gong. Towards zero-shot learning generalization via a cosine distance loss. *Neurocomputing*, 381: 167–176, 2020.

[6] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

[7] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[8] Qunbo Wang, Wenjun Wu, Yongchi Zhao, and Yuzhang Zhuang. Graph active learning for gcn-based zero-shot classification. *Neurocomputing*, 435: 15–25, 2021.

[9] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (9):2251–2265, 2018.

[10] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5542–5551, 2018.

[11] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2021–2030, 2017.

Evaluating GANs for Dataset Augmentation

Francisco Brilhante Fernandes
franciscof@student.dei.uc.pt
Catarina Silva
catarina@dei.uc.pt
Bernardete Ribeiro
bribeiro@dei.uc.pt

Universidade de Coimbra
CISUC - Centro de Informática e Sistemas
FCTUC-DEI - Departamento de Engenharia Informática
Coimbra, Portugal

Abstract

Generative Adversarial Networks (GANS) have recently had a surge of interest for dataset augmentation. Although results seem extremely positive, when data is scarce, there is still slim research in the analysis of their performance, since it is usually hard to perform such evaluation with limited data. In this work we present an analysis of the currently most used indicators to evaluate the quality of the images generated by GANs together with a new combined approach based on nearest neighbour distance to assess how creative GANs are when generating new images. Comparison results on different datasets are also presented.

1 Introduction

Since their introduction in 2014 by I. J. Goodfellow *et al.* [2], the area of Generative Adversarial Networks (GANs) has seen an exponential growth, following the trend in the machine learning field. However, despite the advancements, there is still debate on the best way to evaluate the quality of results produced by each individual GAN in dataset augmentation.

The main objective of this paper is provide an overview of the currently most used indicators to evaluate the quality of the images generated by GANs, namely Fréchet Inception Distance (FID), Inception Score (IS) and Likeness Score (LS). Additionally, a new indicator is proposed to cover the shortcomings of previous methods.

2 Background: GAN analysis

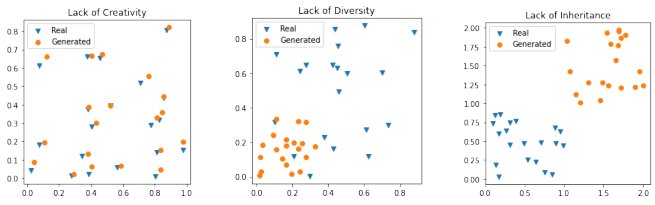
Before diving into any specific method, it is important to establish the aspects that distinguish a good GAN. As stated by S. Guan and M Loew [3], an optimal GAN is one which generates data that follows three characteristics:

-**Creativity**: generated images should be distinct (not copies) from the real ones, and, ideally, bring new information to the existing dataset;

-**Inheritance**: the generated data must follow the same distribution as the original dataset and have the same high-level visual features;

-**Diversity**: each artificial image should obtain some degree of differentiation from the remaining artificial data. An optimal GAN should not produce the same image (or similar images) repeatedly;

Given these aspects, a good evaluation method should be capable of assessing the degree each one is accomplished by a given GAN.



(a) (b) (c)

Figure 1: Representation of extreme scenarios: (a) Lack of Creativity; (b) lack of Diversity; (c) lack of Inheritance

3 Proposed Approach

To understand how each method is evaluated by the three above principles, 5 different tests were conducted using **CIFAR-10** and **Fashion MNIST** datasets. First, the dataset was divided in two groups of equal size and number of samples per class: one representing real data and another simulating data generated by the GAN. Then, the "artificial" group was progressively exposed to the following disturbances:

- **Black Square Insertion**: in each individual artificial image insert a black square located at a random position and width proportional to the severity of the disturbance;
- **Gaussian Noise Insertion**: add Gaussian noise to each image according to the formula $(1-\alpha)X+\alpha N$. X is the image matrix, N represents the Gaussian noise and alpha is the severity of the disturbance. Larger values of α result in more noise being added;
- **Repetition of artificial data**: according to the severity scale, replace part of the artificial dataset by a repeated artificial image;
- **Replacement by real data**: according to the severity scale, each image of the second group had the probability to be replaced by an image of the first group chosen at random;
- **One class convergence**: progressively fill the second data group with images from just one label class.

Each disturbance was evaluated in various degrees of severity (0%, 25%, 50%, 75%, 100%) and multiple samples in cases where the indicator or the modification possessed stochastic properties.

3.1 Inception Score (IS)

Let $x \in G$ be a single generated image. The Inception Score (IS), proposed by T. Salimans *et al.* [5], uses a pretrained inception model - classifier model trained on ImageNet dataset - to compute the conditional label distribution $p(y|x)$ for every generated image. Images with good quality and relevant features from the real data should obtain a label distribution with low entropy. On the other hand, if the generated data is diverse, the label distribution $(p(y))$ over the entire data should have high entropy. IS calculates the Kullback–Leibler (KL) divergence between these two distribution ($p(y|x)$ and $p(y)$) as follows:

$$IS(G) = \exp(E_G[D_{KL}(p(y|x)||p(y))]) \quad (1)$$

In principle, better performance by a certain GAN translates into larger IS values. As stated by A. Borji [1], this indicator does not analyse in any way the real data, therefore, it is not capable of evaluating the similarities between real and artificial data. Another drawback is the need for an image classification network in order to compute label distributions.

Results show that the IS correlated well with the quality of generated images. However, its value did not decrease in cases where the generated dataset was a partial or total copy of the original dataset, showing that it is unfit to evaluate creativity (see Figure 4).

3.2 Fréchet Inception Distance (FID)

FID (Fréchet Inception Distance) is the most commonly used indicator to evaluate the performance of GANs implementations. It was proposed by M. Heussel *et al.* [4] as a direct improvement to the Inception Score by accommodating real data. It uses one of the last layers of the Inception Model to calculate the activation values for both real and fake data. Considering these values to follow multivariate Gaussian distributions, mean and covariance are calculated for the activation values of the real data (m_R, C_R) and generated data (m_G, C_G). Finally, the FID value is presented as the Wasserstein-2 distance between these two Gaussian distributions:

$$FID(G,R) = W_2((m_G, C_G), (m_R, C_R)) \quad (2)$$

While more robust than the previous method, it still suffers from the same problem of not being able to detect whether the generated images are replicas. In fact it scored lower (better) in those situations.

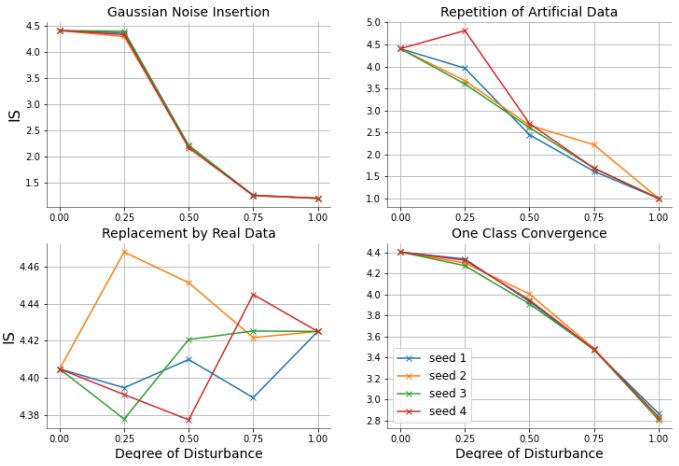


Figure 2: Evolution of IS values computed on Fashion MNIST dataset affected by different modifications. The score was assessed independently 4 times for each level of disturbance.

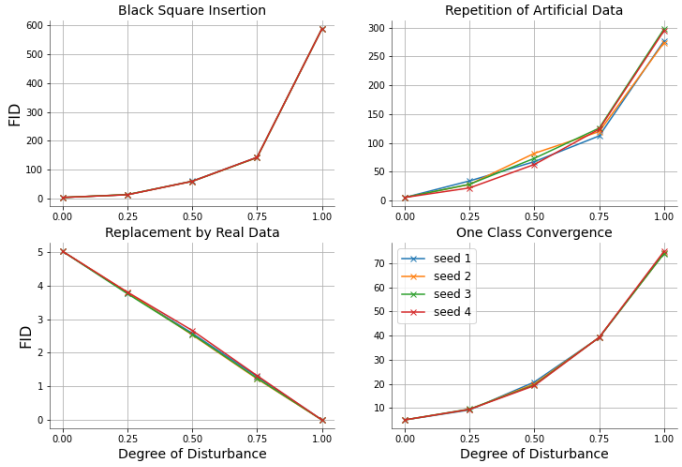


Figure 3: Evolution of FID values on Fashion MNIST with different modifications. Lower values for more similar real and artificial distributions.

3.3 Likeness Score (LS)

The Likeness Score, by S. Guan and M Loew [3], takes the simpler approach of the three methods, by directly analysing the generated and original datasets all together. It starts by defining two different sets - the Intra-Class Distance (ICD) (set of distances between any two points in the same dataset) and the Between-Class Distance (BCD) (set of distances between any two points of different classes/datasets):

$$\{d_x\} = \{||x_i - x_j||^2, x_i, x_j \in X; x_i \neq x_j\} \quad (3)$$

$$\{d_{x,y}\} = \{||x_i - y_j||^2, x_i \in X; y_j \in Y\} \quad (4)$$

First, the ICD sets of generated images (G) and real images (R) (d_g, d_r) and the BCD set ($d_{r,g}$) are computed. Secondly, we use the Kolmogorov-Smirnov distance to compute the similarities between ICDs and BCD (eqs. (5) and (6)). Finally, the Likeness Score is given by the maximum of the two distances (eq. (7)).

$$s_R = KS(\{d_R\}, \{d_{G,R}\}) \quad (5)$$

$$s_G = KS(\{d_G\}, \{d_{G,R}\}) \quad (6)$$

$$LS(G,R) = \max\{s_R, s_G\} \quad (7)$$

While not being as sensitive to the repetition and not accounting for the lack of creativity, LS presented similar results to the previous techniques.

3.4 Combined Approach: Nearest Neighbour Distance

Here, we propose a new method to assess how creative GANs are when generating new images. Despite the robustness of the 3 previous indicators, a common issue was the fact that none of them could detect replicas.

First, for each generated image, is calculated the distance to the nearest real data point. Then, for each real image, is calculated the distance to its closet real neighbour:

$$\{d_{G,R}\} = \{\min\{||x_i - y_j||^2\}, x_i \in G, y_i \in R\} \quad (8)$$

$$\{d_R\} = \{\min\{||y_i - y_j||^2\}, y_i \neq y_j \in R\} \quad (9)$$

After computing these two sets of distances, we apply the Wasserstein distance to obtain the final score:

$$NND(G,R) = W_1(\{d_{G,R}\}, \{d_R\}) \quad (10)$$

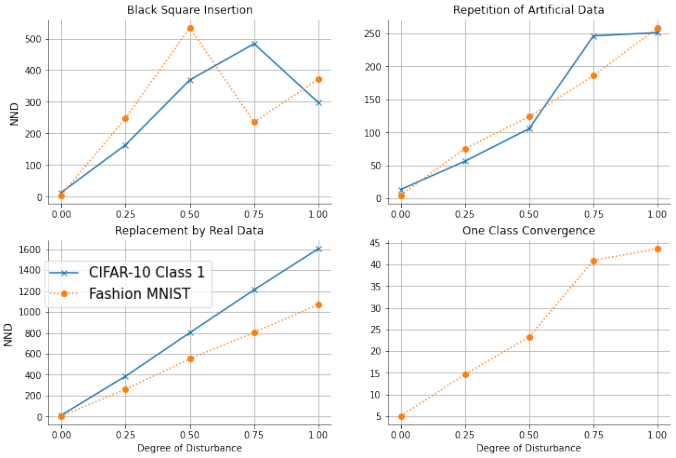


Figure 4: Evolution of the newly proposed NND computed on CIFAR-10 (class=1) and Fashion MNIST datasets with different modifications.

4 Conclusions and Future Work

This paper presented an analysis of the currently most used indicators to evaluate the quality of the images generated by GANs and present comparison results on different datasets.

We proposed a new combined approach, Nearest Neighbour Distance, that assesses how creative GANs are when generating new images by measuring how close each generated image is to the closest real image. Despite being tailored to the creativity principle, the new method performed relatively well in all experiments, on par with the remaining indicators.

Future work is foreseen in observing how the proposed method NND may be improved by taking into consideration k neighbours instead of solely 1 for each data point, in resemblance to the popular k-nearest neighbors algorithm. Furthermore, indicators of the same nature should be accommodated for other types of media generated by GANs, such as audio and video, where current research is scarcer.

Finally, some methods like the NND or Likeness Score may even be applied to other Data Science fields in which new sets of images need to be matched with previously seen data.

References

[1] A. Borji. Pros and cons of gan evaluation measures. *Preprint submitted to Journal of Computer Vision and Image Understanding*, 2018.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *arXiv preprint arXiv:1406.2661v1*, 2014.

[3] S. Guan and M Loew. A novel measure to evaluate generative adversarial networks based on direct analysis of generated images. *arXiv preprint arXiv:2002.12345v4*, 2021.

[4] M. Heussel, H. Ramsauer, T Unterthiner, B Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*, 2017.

[5] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498v1*, 2016.

Real-time pulse rate variability for remote autonomic assessment.

Pedro Constantino[1]
pedroc_24@hotmail.com)
Hugo Plácido da Silva[2]
fhsilva@lx.it.pt
Miguel Constante (PD, MD)[3]
jmiguelconstante@googlemail.com
J. Miguel Sanches[1]
jmrs@tecnico.ulisboa.pt

[1] Institute for Systems and Robotics (ISR), LARSyS, Instituto Superior Técnico, Departamento de Bioengenharia, Universidade de Lisboa
[2] IT - Instituto de Telecomunicações Instituto Superior Técnico Lisbon, PT
[3] Department of Psychiatry, Hospital Beatriz Ângelo - Loures, Portugal

Abstract

Remote medicine is an emerging and important field with the potential to improve patients' health at the distance of a teleconsultation. Here, we propose a novel remote photoplethysmography algorithm suited to extract pulse rate variability in real-time from the face of a patient that is being recorded by a consumer-grade webcam. Because the autonomic nervous system plays a big role in regulating the human heart rate, a remote, real-time pulse rate variability sensor might be of interest for telepsychology and telepsychiatry alike. We test the real-time algorithm with an experiment where both PPG and rPPG are recorded at the same time, from a bluetooth PPG finger sensor and a computer webcam pointed at the face, respectively and where we can see that both signals have similar periodicity, a part from a phase difference.

1 Introduction

For several years physicians have monitored heart rhythms through auscultation and have noted that beat-to-beat times shift depending on age, illness and psychological state [1]. Both electrocardiography and photoplethysmography are used to access cardiovascular signals (see Fig. 1).

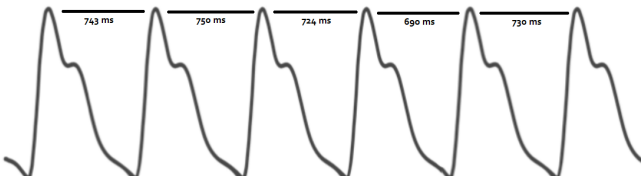


Figure 1: Five time differences (in milliseconds) between six pulses via photoplethysmography. This figure is not in scale.

But, while the electrocardiogram is a measure of electrical activity directly related to the contractions of muscular heart, the photopletysmographyic signal is an optical measure that captures the amount of blood coming and going from a given tissue, and thus only indirectly it captures the heart's beating. From electrocardiography, heart rate (HR) is defined as the number of heartbeats per minute and heart rate variability (HRV) concerns the fluctuation in the time intervals between adjacent heartbeats. Similarly, we can define pulse rate (PR) and pulse rate variability (PRV) in the context of photoplethysmography.

In 2018, remote photoplethysmography (rPPG) was reportedly [5] the most popular name for a technique that can also be referred to as contactless PPG, camera-based PPG or imaging PPG. Aside from a source of light, the only component needed is a camera (e.g. low-cost webcam, mobile phone camera), which makes this technique really promising for the telemedicine context.

1.1 Heartbeat and autonomic regulation

Two types of cardiac muscle cells generate the heartbeat: (1) contractile cells produce strong contractions that cause the heart chamber to shrink and propel blood, and (2) specialized noncontractile muscle cells of the conducting system control modulate contractile cells. Contractile muscle cells, which comprise the majority of cardiac muscle cells, are activated by external action potentials, similarly to skeletal muscle. On the other side, noncontractile muscle cells are less in number and organized as a network made up of two types of cells: nodal cells and conducting cells. Nodal cells are autorhythmic, i.e. they contract on their own, without neural or hormonal stimulation, and generate the pacemaker potentials responsible for initiating the muscular heartbeat. They are located at the Sinoatrial (SA) and Atrioventricular (AV) nodes. However, nodal cells from the SA node naturally depolarize faster, 70–80 action potentials per minute, than those in the AV node, 40–60 action potentials per minute, being the effective pacemaker cells in the heart. [4]

Although the SA node spontaneously generates the normal heartbeat cardiac rhythm, autonomic motor neurons, circulating hormones and ions can influence the inter-beat interval and magnitude of the myocardial contraction [34]. More specifically, the cardiovascular center, located in the brain stem, integrates sensory information from various bodily receptors and responds through sympathetic and parasympathetic motor neurons (and endocrine systems), adjusting the HR continuously [3]. See Fig. 2.

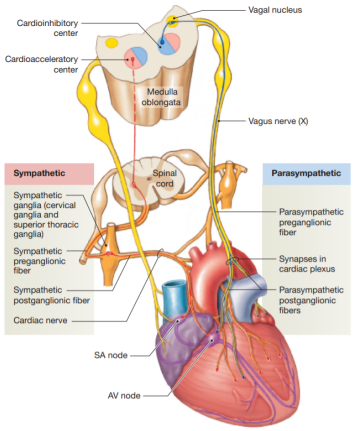


Figure 2: Autonomic innervation of the heart. Adapted from (Martini, 2018).

Cardiac sympathetic nerves target the SA node, AV node, and the bulk of the myocardium and trigger norepinephrine and epinephrine release and binding to beta-adrenergic (b1) receptors located on cardiac muscle fibers, speeding up spontaneous depolarization in the SA and AV nodes (increasing HR) [2]. The parasympathetic vagus (X) nerves also innervate the SA node, AV node, and atrial cardiac muscle and trigger acetylcholine release and binding to muscarinic receptors, decreasing the rate of spontaneous depolarization in the SA and AV nodes (slowing HR) [6].

2 Method

The proposed method takes as input a timestamped stream of video and it is assumed that all frames of the video contain a face.

First, for every video frame, $frame[i]$, an average of the RGB channels over a predicted facial skin region of interest, $avg\_rgb[i]$, is produced. This can be accomplished with face detection, facial landmarks prediction and ROI selection, as seen in fig. 3. Here, the regions comprising the cheeks, the nose and the forehead are considered, while the regions for beard and eyes are removed. Since following algorithms assume that samples are evenly distributed in time, resampling the asynchronous RGB signal is needed.

This work was supported by Portuguese funds through FCT (Fundação para a Ciência e Tecnologia) through the projects reference UIDP/50009/2020 (Programático) and through the reference UID/EEA/50009/2019, LARSyS - FCT Plurianual funding 2020-2023.

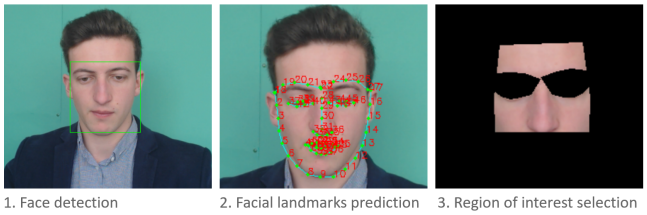


Figure 3: The first three steps of the algorithm.

After that, we must estimate a blood volume pulse (BVP) signal, $bvp[t]$, from the collected RGB signal, $avg\_rgb[i]$. For that we apply the POS method to transform the skin RGB signal into a BVP signal [8]. A bandpass filter is further applied to clean the signal for peak detection. Cut off frequencies were set as [0.8, 2.5] Hz, since these frequencies correspond to a normal human heart rate range of 48 to 150 bpm. See fig. 4.

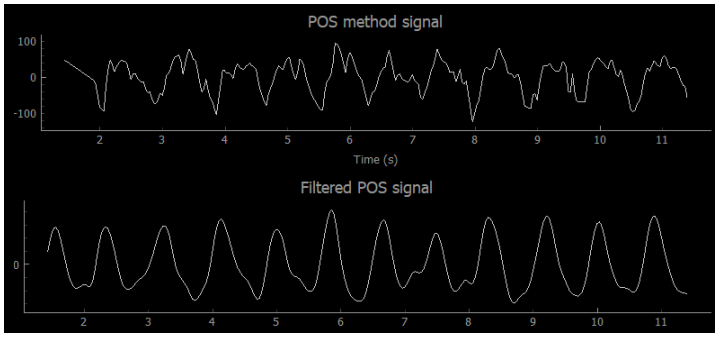


Figure 4: On the top, we can see the blood volume pulse obtained from the RGb signal via the POS method and, below, we see the filtered BVP.

Lastly, PR and its variability are estimated via the filtered BVP signal. Peak detection is firstly applied to the BVP signal in order to detect true heartbeats and, from the peaks, we can compute PP intervals. See fig. 5.

We extract PP intervals as the difference between pairs of consecutive peaks.

$$PPinterval_i = IBI_i = PeakTime_i - PeakTime_{i-1} \quad (1)$$

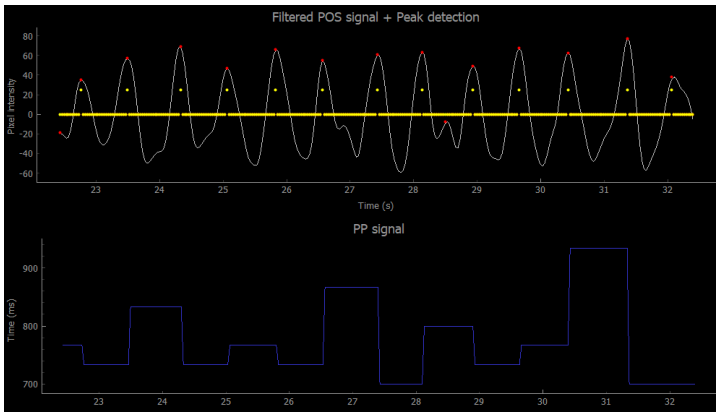


Figure 5: On the top, we can see the peak detection; below, we see the PP signal, or IBI signal.

From the PP intervals, we can provide estimations of PR and time-domain measures of PRV. For a full review on PRV metrics see [7].

3 Results and Discussion

To test the proposed algorithm, a standard grade laptop (MSI GF63 8RD) runs the whole rPPG pipeline in real-time using the embedded webcam as video input in one thread, and, on another thread, it acquires the traditional PPG signal from a pulse oximeter finger clip sensor, i.e. the ground truth signal. The PPG acquisition is mediated through a BITalino board, which stores the data at a constant rate and sends it via bluetooth to the laptop.

In the end, we can display both signals at the same time and confirm that the rPPG signal follows the PPG signal closely (see Fig. 6), though we can see that the two signals are not perfectly aligned. This delay might be related with: 1) PPG thread starting acquisition first than the webcam thread, or vice versa and 2) the amount of time blood takes to travel from the heart to the face is different from the time it takes travelling from the heart to the finger, and their also target of regulation by the circulatory system control mechanisms. Anyway, we can see that for every PPG-sensor pulse we can count a corresponding delayed rPPG-sensor pulse, confirming the ability of the proposed real-time algorithm to capture pulse rate variability, just like traditional PPG can do.

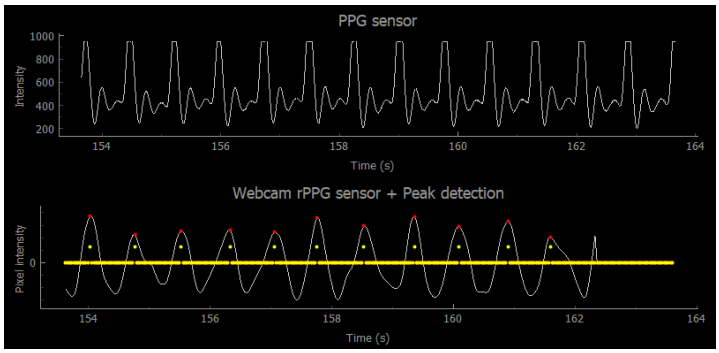


Figure 6: Real-time comparison of the photoplethysmography signal obtained through the finger clip sensor (upper panel) and the remote photoplethysmography signal obtained from the webcam video.

4 Conclusions

Accomplishing real-time pulse rate variability means that we can inspect the raw signal and PRV features during recording, which allows to identify artifacts, make sense of the values and overall have control over the precision of the process. Plus, a real-time rPPG algorithm ensures the doctor has some control over the quality of the measurement. The real-time display not only allows him to search for a sweet spot in terms of patient positioning and ambient lighting, but also to control overall quality of the record. This work is relevant because telemedicine is an emerging, cheaper form of providing health services and reliable tools must be developed to support doctors in making decisions within the remote context.

References

[1] Gary G Berntson, J Thomas Bigger Jr, Dwain L Eckberg, Paul Grossman, Peter G Kaufmann, Marek Malik, Haikady N Nagaraja, Stephen W Porges, J Philip Saul, Peter H Stone, et al. Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology*, 34(6):623–648, 1997.

[2] HF Brown, Dario DiFrancesco, and SJ Noble. How does adrenaline accelerate the heart? *Nature*, 280(5719):235–236, 1979.

[3] BCB Fred Shaffer PhD and John Venner MAE. Heart rate variability anatomy and physiology. *Biofeedback (Online)*, 41(1):13, 2013.

[4] Frederic H. Martini, Judi Lindsley Nath, Edwin F. Bartholomew, and William C. Ober. *Fundamentals of anatomy & physiology*. Pearson, 2018.

[5] Philipp V Rouast, Marc TP Adam, Raymond Chiong, David Cornforth, and Ewa Lux. Remote heart rate measurement using low-cost rgb face video: a technical literature review. *Frontiers of Computer Science*, 12(5):858–872, 2018.

[6] Bert Sakmann, A Noma, and W Trautwein. Acetylcholine activation of single muscarinic k+ channels in isolated pacemaker cells of the mammalian heart. *Nature*, 303(5914):250–253, 1983.

[7] Fred Shaffer and James P Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, 5:258, 2017.

[8] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016.

Organization of Information in Feed-Forward Neural Networks

Ricardo Coke
itsricoke@gmail.com

ECT-School of Science and Technology
University of Trás-os-Montes e Alto Douro UTAD
Vila Real, Portugal

Paulo Salgado
psal@utad.pt

ECT-School of Science and Technology
University of Trás-os-Montes e Alto Douro UTAD
Vila Real, Portugal

Abstract

The neural network complexity is intrinsically associated with parameters such as the number of hidden layers, activation function, which influence how flexible are the network pattern classifications, allowing for easier classifications in higher spatial dimensions at the cost of more resources, time and interpretability when computing the problem. In order to solve the increased computational needs and lack of interpretability, new solutions and structure modifications to the typical networks have been researched and implemented. This article presents an algorithm based on a newly developed complementary NN model. A classification problem was proposed and solved by a standard NN and the newly developed model presenting similiar accuracy's resulting in a working NN which can be further developed to achieve higher degrees of interpretability manipulating its complementary feature.

1 Introduction

The action potential presented in the neurons of the human brain has different pulse patterns with distinct distances between pulses [3]. In order for the same to happen in artificial neural networks, complex numbers or modifications to the default structure can be introduced to mimic the signal phase and amplitude [2]. This concept led to the use of complex numbers resulting in Complex-Valued NN's or Multi-Valued Neurons resulting in Multi-Valued NN's, which are excellent function approximators being able to preserve two signals [1].

2 Methods

The experiment data used involved a non-linear function that the NN had to approximate with a high degree of precision. The chosen function represented four semi-circles that were divided by the y-axis, having two in the left side and two in the right side represented in Fig 1. When the interior semi-circle belonged to class 1, the exterior semi-circle had to belong to the other class, class 2. In order to use dual-valued inputs to the NN a 2 by 1 matrix was used. This matrix corresponds to the same amount of information present in the one-valued neuron, where each input only corresponded to one value. In other words, the sum x1 (first component) and x2 (second component) will result in the complete x (input). Distinctly from a standard network, there is a new parameter associated to the inputs, which is the alpha value and varies between 0 and 1, representing how much of the information will be stored in the first and second component. This makes the transfer function of the dual-valued neuron:

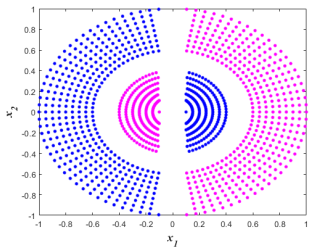


Figure 1: Classification Problem

$$y_i = \sum_{i=0}^{m} \alpha W \omega_i(n) x_i(n) + b_i * \alpha b \tag{1}$$

The output of the neural network will also be composed of two parts (first and second component), with $y = y_{i,1} + y_{i,2}$. Regarding each of the output singularly ($y_{i,1}$ and $y_{i,2}$), they represent what's obtained after passing all the information through the activation function:

$$y_n = \varphi_n(\alpha_1, \alpha_2) \quad \text{where} \quad n = 1,2; \tag{2}$$

2.1 Weight Calculation

The total weight will be obtained through the standard NN transfer function.

$$y_i = \sum_{i=0}^{m} \omega_i(n) x_i(n) + b_i \tag{3}$$

This total value can be unfolded in two components: $W_x = W_{x1} + W_{x2}$. By calculating the total weight and one of the component's weight (by applying equation 1), the following equation can be applied to obtain the remaining component, $W_{x2} = W_x - W_{x1}$

2.2 Activation Function

The activation function chosen for the hidden layers was the hyperbolic tangent with a slight modification. This new sigmoidal bipolar "TanSig" activation function is approximated by the Fourier Series at $2m-1$ terms, with a period of 16 ($Ts = 16$), and with its fundamental frequency given by $\omega_0 = \frac{2\pi}{Ts} = \frac{\pi}{8}$

Since we need two activation functions, to the first and second components, that when summed represent an output exactly like the one obtained from a one-valued neuron, we'll make use of functions that complement each other.

$$\begin{aligned} \varphi_1(a_1, a_2) &\approx \sum_{i=1}^{m} \sigma_i sin(\omega_i a_1) cos(\omega_i a_2) \\ \varphi_2(a_1, a_2) &\approx \sum_{i=1}^{m} \sigma_i cos(\omega_i a_1) sin(\omega_i a_2) \end{aligned} \tag{4}$$

With $\omega = (2i-1)\,\omega_0$. This additive property can be verified by:

$$\varphi_1(a_1, a_2) + \varphi_2(a_1, a_2) \approx \sum_{i=1}^{m} \sigma_i \, sin(\omega_i (a_1 + a_2)) \approx tansig(a) \tag{5}$$

2.2.1 Dual-Valued Feed-Forward

The first layer output (y), which is composed of two components y_1 and y_2, is given by: $Y = \varphi(W_1 \times x + b_1)$, where x represents the inputs. For the first and second components output calculation, the following equations will be used:

$$\begin{aligned} Y_1 &= \varphi\left[(W_1 \times \alpha W_1) \times x + b_1 \times \alpha b_1\right] \\ Y_2 &= Y - Y_1 \end{aligned} \tag{6}$$

For the remainder layers (i), in order to use the output obtained from the previous layer as the input:

$$\begin{aligned} Y_i &= \varphi\,(W_i \times Y_{(i-1)} + b_i) \\ Y_{i,1} &= \varphi\left[(W_i \times \alpha W_i) \times Y_{(i-1,1)} + b_i \times \alpha b_i\right] \\ Y_{i,2} &= Y_i - Y_{(i,1)} \end{aligned} \tag{7}$$

2.2.2 Backward Computation

To compute the local gradients and weight adjustment since both components are complementary, if we adjust the first component, compute the entire model as if it was a one-valued network, and subtract the entire

model by the first component, the result is the second component. The cost function will be defined by the gap between the output obtained and the desired output:

$$C = \frac{2}{N}(y_k - d_k) \tag{8}$$

The error of each iteration will be given by the sum of all values provided by the cost function in the given iteration.

2.2.3 Error Propagation

The method of Ordinary Least Squares will be used to retro-propagate the error. When in presence of the "Tansig" activation function in the output layer, we'll need the activation function, equation 4, derivative in order to the real term (α_1):

$$\begin{aligned} \frac{\partial\, \varphi_1(\vec{a})}{\partial \alpha_1} &== \sum_{i=1}^{m} \sigma_i\ \omega_i\ cos(\omega_i(\alpha_1 - \alpha_2)) \\ \frac{\partial \alpha_2}{\alpha_1} &= \frac{d(\alpha - \alpha_1)}{\partial \alpha_1} = -1 \end{aligned} \tag{9}$$

Where σ_i represents the coefficients of the Fourier Series.
To obtain the input derivative, having the sub model derivatives and computing them with the output obtained from the forward computation (y_1, y_2), the real term derivatives (α_1) are multiplied by the gradient (derivative of the cost function):

$$\delta_{nc} = F' \times \left[\frac{2 \times (y_i - d_k)}{N}\right] \tag{10}$$

Where $[\frac{2\times(y_i - d_k)}{N}]$ is the gradient. For the hidden layers:

$$\delta_i = F' \times \left(\omega_{(i+1)} \times \alpha W_{(i+1)} \times \delta_{(i+1)}\right) \tag{11}$$

2.2.4 Weight Adjustment

In order to maintain higher stability during the training phase, a stability heuristic using the learning rate was implemented.

$$\eta_{iu} = \eta \times \left(0.5 + 0.5 \times \frac{(i-1)}{(nc-1)}\right) \tag{12}$$

Starting from the output layer and moving backwards to the second layer, each iteration, the values of the alpha variables (weights and bias) will be updated.

$$\alpha W' = \sum(\delta_i \times (\omega_i \times y_{(i-1,1)})) \tag{13}$$

The descent gradient is then applied, updating the value of αW for the current iteration:

$$\alpha W_i = \alpha W_i - \eta_{iu} \times \partial \alpha W' \tag{14}$$

Regarding the adjustment of the alpha bias variable, a higher learning rate is used since in our optic the bias offset should be adjusted more than the weights when in an attempt to separate the network information:

$$\alpha b_i = \alpha b_i - 4 \times \eta_{iu} \times \sum(\delta_{(i,2)} \times b_i) \tag{15}$$

For the first layer the weights and bias used will multiply by the network input (x), instead of the previous layer output ($y_{(i-1)}$), as done for the hidden layers.

$$\alpha W_1' = \sum(\delta_i \times (\omega_i \times x_{(1,1)})) \tag{16}$$

$$\alpha W_1 = \alpha W_i - \eta \times 0.5 \times \alpha W_1' \tag{17}$$

$$\alpha b_1 = \alpha b_1 - 4\eta \times \sum(\delta_{(1,2)} \times b_1) \tag{18}$$

2.2.5 Stopping Criteria

The stopping criteria will be the minimization of the error until is inferior to the desired precision ξ. Considering $(d_{k,1} + d_{k,2}) = (y_{k,1} + y_{k,2} + e_k)$ and using the method of minimizing the sum of the quadratic error applied to two outputs we have:

$$E = \frac{1}{N}\sum_{k=1}^{N} 2(e_{k,1} - e_k)e_{k,1} + e_k^2 \tag{19}$$

Where $d_{k,1} = \delta_k d_k$, and $d_{k,2} = (1 - \delta_k)d_k$.

3 Results

3.1 Mean Squared Error

The Mean Squared Error (MSE) results provided by the standard NN was 0.0003 while the DVNN first and second component after the learning process, were 0.00433 and 0.00346 (respectively), which represents a very close approximation of the desired values by the network.

3.2 Confusion Matrix and associated metrics

The results of the Confusion Matrix (CM) in the standard NN were perfect with no misclassified classes resulting in Precisions, Recall, F1 and Specificity of 1 and having an error of 0. When in presence of a Multiclass CM, its construction relies on the comparison of the predicted class with the known true class . The confusion matrix resultant of the training data is represented in Figure 2 for the first component and second component.

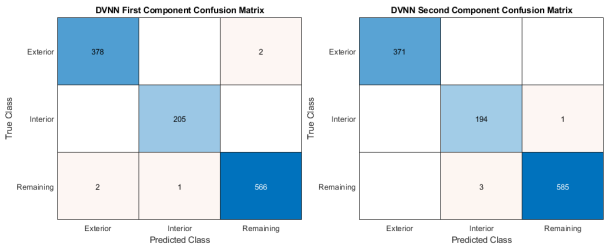


Figure 2: DVNN first and second component Confusion Matrix

The following results were provided by the "metrics" function developed, regarding the first and second component of the DVNN (average):

Metric	Precision	Recall	F1	Specificity	Error
Exterior	0,9948	0,9948	0,9873	0,9972	0,0022
Interior	0,9957	0,9948	0,9961	0,9974	0,0013
Remaining	0,9976	1,0000	0,9988	0,9995	0,0004
Mean	0,9950	0,9966	0,9958	0,9980	0,0013

Finally the Standard NN accuracy was 1 while the new model presented 0.9961 average accuracy (first and second component) .

4 Conclusions

The study presented the methods and implementation needed for the development of a DVNN, and the respective results given a classification problem. This results showed a DVNN capable of learning, storing the information in a complementary manner with high peformance in typical ML performance metrics as demonstrated in Chapter 3 which rival the standard NN model.The complementary channels of the DVNN also allow for further implementations and experiments that might allow for easier transferability of information between networks and interpretability, which are advantages when compared to a standard NN.

References

[1] Igor Aizenberg and Claudio Moraga. Multilayer feedforward neural network based on multi-valued neurons (mlmvn) and a backpropagation learning algorithm. *Soft Computing*, 11(2):169–183, 2007.

[2] Md Faijul Amin and Kazuyuki Murase. Single-layered complex-valued neural network for real-valued classification problems. *Neurocomputing*, 72(4-6):945–955, 2009.

[3] Tohru Nitta. An extension of the back-propagation algorithm to complex numbers. *Neural Networks*, 10(8):1391–1415, 1997.

Adaptive body interface to control devices using KNX protocol

Jedid-jah Dorneles dos Santos[1]
jedid.santos@gmail.com
Ivo Manual Valadas Marques Martins[2]
immartin@ualg.pt
João Miguel Fernandes Rodrigues[3]
jrodrig@ualg.pt

[1]ISE, Universidade do Algarve, Portugal
[2]INESC-ID & ISE, Universidade do Algarve, Portugal
[3]LARSyS & ISE, Universidade do Algarve, Portugal

Abstract

The use of automation equipment in our daily lives, for our daily activities and assistance in our work, has been growing exponentially. Every day people want a comfortable environment for leisure or for help with their work activities, including, of course, people with different types of disabilities. For this, it is important to develop interfaces that can be adaptable to each user needs and wills. This article presents a framework that integrates human actions: gestures and/or poses to activate automation devices that communicate by KNX protocol. The pose detection algorithm is used to learn and detect different gestures/poses, where each gesture or group of gestures integrated with the KNX protocol allows easy and universal communication with various types of existing automation devices. The results show that the framework can adapt smoothly to each user and perform the interaction between the user and the equipment.

1 Introduction

The development of technological platforms and the effects of their proliferation on home, business, and social enterprises is one hot research topic. Special attention must be given to adaptive interfaces [1], that is, interfaces that can adapt in real-time for each user. More and more people want a comfortable environment for leisure or work, something practical, sustainable, and safe. Achieving this level of automation today is a real challenge, once it involves a huge amount of "wiring" to implement the installations, from sensors and actuators to control and monitoring centres. Of course, there are several solutions to minimize this, one of them is the use of equipment that employs the KNX protocol for communication between devices [2].

KNX is a technology for home automation, which allows the control of different devices such as lighting, shutters, security, energy management, HVAC systems, etc. KNX is also the only open standard for residential and building control [2]. With the use of the KNX protocol, it is not necessary to use isolated sensors and actuators, as this technology allows communication over the internet (IP) with all these devices simultaneously.

This paper focuses on how to command devices that are controlled by the KNX protocol using human actions, gestures or body movements defined by the user. It is the continuation of an initial proof-of-concept [3], which presents an Adaptive Control Devices Framework (ACDf) that uses predefined person actions (movements) to control devices using the KNX standard. The ACDf2.0 uses MediaPipe [4] library instead of the previous gesture detection framework, GluonCV [5], and uses a Neural Network to learn the human actions, i.e, user gestures or actions (body, head or hand) is a way to adapt each command (interface) to each different user. The main contribution of the paper is the adaptive user interface that is implemented in conjunction with the KNX protocol to command devices.

The ACDf2.0 is divided into five main modules: (a) Communication with the devices (via KNX); (b) Detection of human movements; (c) Association between human movements, actions, and device functionalities; (d) Learning personalized actions - development of the adaptative interface; (e) Store and share actions over the internet (cloud). Modules (a) to (c) were detailed in [3]. Here we focus on the remaining modules, despite as already mentioned, we use MediaPipe instead of GluonCV for de detection of human movements (module (b)).

2 ACDf2.0

Despite the above mentioned, a brief explanation of the first three modules is necessary: (a) The communication with devices is done through the standard KNX protocol through an open-source library[1]. This library is implemented in Python, where it is possible to connect through IP the KNX devices, being possible to activate their functionalities. (b) The motion detection initially used the GluonCV toolkit [5], it is a deep learning tool for Computer Vision that works in Python where we can find classification algorithms, object detection, pose estimation, among others. However, during the development of this framework and for the modules that will be presented, the body keypoint detection tool was changed to MediaPipe [4], as it returns better results. MediaPipe is also an open-source cross-platform, customizable ML solution for live and streaming media (as mentioned in their site[2]); (c) In ACDf the association of human movements with the features of the devices was hard-coded, i.e., was only possible to do a predefined group of movements to perform certain actions. Actions are, for instance: (i) *start* the interaction with devices, (ii) *next* and (iii) *previous* that allows, for example, to navigate between devices, when there is more than one device to be controlled. Activate or disconnect devices, (iv) *turn on* and (v) *turn off*, and if the device has the functionality to increase or decrease a certain parameter, the actions (vi) *increase* and (vii) *decrease*. Details can be seen in [3].

Module (d) consists of the development of the adaptive interface, where each user can, if desired, teach the system their movement(s) to associate to each action. The idea is always to minimize and simplify as much as possible the number of actions. This is because a young person can, for example, make moves that an older person cannot do. Finally, module (e) consists of storing in the cloud all the correspondence of movements and functionalities of the devices so that the same user can be anywhere/ambient. This means that after a set of actions is learned by the framework (in module (d)) and associated with that user (in module (c)), these actions can be used in any place or environment (in the world), without any kind of configuration (extra).

For the framework to be adaptable and learn new gestures, module (d), is used a neural network architecture called Long Short-Term Memory (LSTM) which are a type of recurrent neural network capable of learning order dependence in sequence prediction problems, that is, it is a recurrent neural network in deep learning that was specifically developed for the use of treatment of sequence prediction problems [6].

This requires several steps: (d.1) The first step is to extract all the keypoints of the pose, face, left hand, and right hand using the MediaPipe library [4] (done in module (b)), since we do not know if the user will use a "body movement" or a hand movement or a combination of various movements, including facial movements. In summary, this library has 1662 keypoints being distributed by the user's body and that will be extracted for network training. (d.2) These keypoints after-acquired are flattened into a single vector, which will be used later to feed the network. It is important to emphasize at this point that every time a keypoint is not detected, its values are set to 0.

(d.3) The next step is to collect a set of frames (movies) that represent each movement, to teach the different movements to the network. In our case, we are not interested in the "movies", but the keypoints extracted by MediaPipe. We must acquire a movement, and for these movements, we must acquire n sequences for that movement (more sequences, more data, in principle better results, in the end, this trade-off between acquiring on-the-fly more data to achieve better results and the tedious for the user), and the length of each sequence, in frames is ls. This means that we have for each action $nfa = ns \times ls$ (number of) frames. In our case, $na = 6$, $ns = 10$ and $ls = 30$.

(d.4) Now that we have the dataset of the movements, we intend to use, the next step is to partition the data into two: (i) training data and (ii) test data, giving them a label that we'll call "action 1" to "action 6" which is the number of ACDf actions. The partitioning of the data is being 95% for training, 5% for testing (validation). The reason for these values is that our dataset is very small, with larger datasets the partition should be around 60% for training and 40% for validation and testing.

(d.5) The next step is to train the network. Our network is a standard 6 layers network where the first three layers are LSTM [6], which allows a temporal component for learning movements. The first layer has 64

[1] XKNX: *Asynchronous Python Library for KNX*, https://xknx.io/

[2] MediPipe: https://mediapipe.dev/

neurons/units, the second 128, and the third 64. After the first 3 layers are followed by 3 fully connected layers, with 64, 32, and the last layer with 6 neurons. The number of neurons in this layer is equal to the number of moves we want to learn, in our case 6.

(d.5) It is now possible to save the model for the movements. The final step in ACDf2.0 is to load them (model) every time we start our application and make the inference to know what movement is being performed by the user (instead of the predefined ones done in ACDf).

Going back to the storage of actions in the cloud, module (e), an Application Programming Interface (API) provided by Google Drive [7] is used, compatible with Python (all modules were done using Python), so that it is possible to upload in a personal account the files that the user has locally on the device anywhere the ACDf2.0 is used. This way, the user can perform body/gesture training for an installation with KNX devices at home, for example, and can load these same movements and used them in his office or in any other environment that has the associated framework. The process for this storage is carried out in a simple way where an authorization (token) and an account on google drive is required in a way that it is possible to have the access release and make changes to the destination folder.

3 Tests

To verify the functionality of the ACDf2.0 with the movements taught and storage in the cloud, tests were carried out following the same methodology presented in [3]. Using 6 users, divided into groups of 2 in 3 different scenarios (laboratory, home, and office). Each user performed the process described so that the framework could learn the new gestures they thought more appropriate for the action (Fig. 1 top rows). After the training, activation and interaction with the KNX installation were carried out (Fig 1 bottom) and the files were uploaded to the corresponding cloud of the user. In total, 10,800 frames were collected, where 6 users × 6 actions × 10 videos × 30 frames (each video).

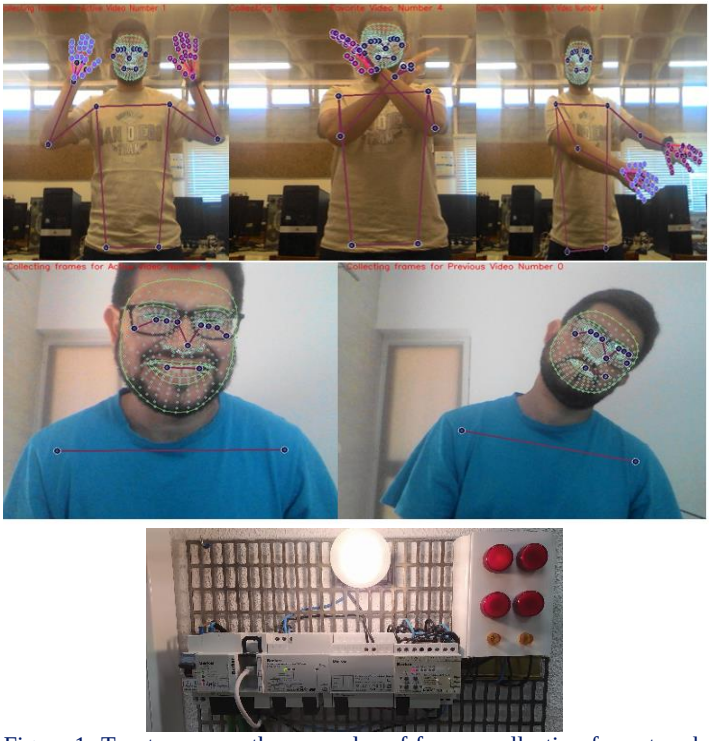


Figure 1: Top two rows, the examples of frames collection for network training for gesture/pose recognition using MediaPipe [4]; bottom row an installation of a KNX router and devices connected to a lamp.

Table 1 shows the percentage of correct results performed with the movements trained by the users. In terms of results, the tests show lower results (average of 61.7%) than the tests performed in ACDf [3], but this was expected because now the movements are being trained, by a small number of repetitions, and in addition, the user is not doing aways the same “expected correct movement”.

Nevertheless, this does not make the results unsatisfactory, because, in the same way, that the communication with the pre-defined movements was carried out, now it happens with the movements trained by the user.

The training was also carried out, by considering movements with the face (smiling, lateral movement of the head, etc.) and it was stated that without more specific treatment for this situation, the results are poor, as it is necessary to be very close to the camera to detect the action, once MediaPipe detects a large number of keypoints and as the distance increases, it becomes more difficult to differentiate whether a particular move was performed. In relation to distance, it was also verified that when the user starts to get away from the camera that is capturing the results also do not have a high percentage of correct answers since the keypoints end up overlapping and with this the training is affected, and may not recognize the movement or recognize in an incorrect way.

Related to cloud storage, no problems were found in your application, only when the user enters incorrect credentials it is not possible to perform the storage.

Table 1: Table with the percentage of correct answers obtained during tests the ACDf2.0.

Action	Accuracy (%)
Next	60,0
Previous	62,5
Turn-on/Turn-off	64,0
Increase	65,2
Decrease	62,0
Favorite	56,6

4 Conclusions

It was proposed the continuation ACDf framework, capable of adapting to each user and performing interactions/activations in automation devices with KNX protocol – ACDf2.0.

The results obtained validated that the framework works as expected, through a group of users, collections of different movements were carried out where the framework was trained using a neural network, these movements replaced the previously defined ACDf movements. With these movements, interactions/activations were successfully carried out in the automation installation. The files were also successfully uploaded with the movements trained so that it was possible to access these movements in different environments where the framework was installed.

Despite the framework showed good results for the proposed situation and is still open to different improvements in future work, such as the mentioned incorrect learning of movements when using the head.

References

[1] Rodrigues, J.M.F., Pereira, J.A.R., Sardo, J.D.P., Freitas, M.A.G., Cardoso, P.J.S., Gomes, M., Bica, P. (2017) *Adaptive Card Design UI Implementation for an Augmented Reality Museum Application*, In M. Antona and C. Stephanidis (Eds.): Universal Access in Human-Computer Interaction 2017, Part I, LNCS 10277, pp. 433–443 (2017) DOI: 10.1007/978-3-319-58706-6_35.

[2] KNX: *KNX Smart Home and Building Solutions*. Global. Secure. Connected, https://www.knx.org/knx-en/for-your-home/, last accessed 2021/09/10.

[3] Santos J., Martins I., Rodrigues J.M.F (2021) *Framework for Controlling KNX Devices Based on Gestures*. In: Antona M., Stephanidis C. (eds) Universal Access in Human-Computer Interaction. Access to Media, Learning, and Assistive Environments. LNCS 12769. Springer, Cham. https://doi.org/10.1007/978-3-030-78095-1_37.

[4] Grishchenko, Ivan., Bazarevsky, Valentin. (2020). *MediaPipe Holistic — Simultaneous Face, Hand and Pose Prediction, on Device*. (online on: https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html, last accessed 2021/09/13).

[5] GluonCV, *State-of-the-art Deep Learning Algorithms in Computer Vision* https://cv.gluon.ai/, last accessed 2021/09/13.

[6] Sherstinsky, A. (2020). *Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network*. Physica D: Nonlinear Phenomena, 404, 132306

[7] Google Drive. (2021). *Introduction to Google Drive API*. (online on https://developers.google.com/drive/api/v3/about-sdk, last accessed 2021/09/10)

Biometric identification and authentication based on electrocardiogram

Bruna Alves
bruna.alves@ua.pt
Raquel Sebastião
raquel.sebastiao@ua.pt

Department of Physics, University of Aveiro, Portugal

Institute of Electronics and Informatics Engineering of Aveiro & Department of Electronics, Telecommunications and Informatics, University of Aveiro, Portugal

Abstract

Biometric systems have gained enormous importance due to their high efficacy and safety. In this context, the focus of this work is the biometric identification and authentication based on the electrocardiogram (ECG) signal. For that, two different ECG-derived models are used and assessed - cardiac cycles and scalograms - comparing the distance (using Euclidean or Manhattan distances) between the test and training sets. The number of models used (10 or 20, for each set) is also evaluated. For the identification task is used the 1-nearest neighbour classifier. For a successful authentication, more than 50% of the distances between the test and training models, of the same individual, must be less than a personalized threshold (obtained for each individual). The identification and authentication rates obtained were 100% when using cardiac cycles. With this signal, a minimum imposter rate of 13.7% is obtained when using the Euclidean distance and 10 cycles. Using scalograms, identification and authentication rates of 90% and 95% were obtained, respectively, with an imposter rate of 12.4%.

1 Introduction

Diverse economic sectors, such as finances, health services, transportation, entertainment, and law enforcement, are examples of access services that require a person identification.

Traditional systems are based on individual representation through tokens, smart cards, or passwords (among others) which can easily be stolen or copy. Moreover, usually password-access requires that passwords should be long and unique, which are difficult to be memorized by users, ending up in easy-to-guess passwords that are commonly reused to grant the same user access to different systems or devices.

Biometric systems use physiological and behavioural characteristics (named biomarkers) to identify or authenticate people. Nowadays, the most common biometric systems are based on facial recognition or fingerprint scanning [4]. However, these methods are not ideal. The use of some objects, *e.g.* glasses, or the alteration of a particular characteristic, such as grown beard or even a scar, can prevent the system to identify or authenticate an individual.

Biometrics based on electrocardiogram (ECG) have been researched over recent years. This physiological signal fulfils the requests to be used as a biomarker: it is universal as it is present for all individuals; it is unique for each person; it is easy to measure using appropriate equipment; it is always available to measure as long as the person is alive and it is fraud-resistant as it is difficult to fake [4].

The main goal of this work is the biometric identification or authentication based on ECG signals. For this purpose, two models were used: cardiac cycles and scalograms. Therefore, a comparison between the two models in both biometric tasks is established.

2 State of Art

Recently several biometric identification and authentication works based the ECG have emerged [1, 2, 3, 6]. Lourenço *et al*. [3] proposed a biometric system based on ECG signals recorded at the fingers. The authors collected signals from 16 subjects. The criterion used for identification was the minimum Euclidean distance between the test and the training sets (nearest neighbour classifier) and the individual is authenticated if the Euclidean distance between test and training sets is inferior to a given threshold. They reported a 94.3% recognition rate in identification and an equal error rate of 13.0% in authentication tasks. To improve the authentication performance, the authors experimented a user-tuned threshold selection method. After applying the user-tuned threshold selection method, the equal error rate improved to 10.1%.

Lee and Kwak [2] used two databases to develop an authentication system: the Chosun University ECG Database (CU-ECG DB), which was built by the authors, and the Physikalisch-Technische Bundesanstalt ECG database (PTB-ECG Database), which is a public database. The CU-ECG has signals from 100 individuals (11 females and 89 males aged between 23 and 34 years old). The PTB-ECG has recordings from 290 individuals (females and males). The authors combined a robust neuronal network (REECGNet), for feature extraction, with a support vector machine (SVM) classifier. They used scalograms, which are time-frequency representations of the ECG signal, as input of the neuronal network. A performance of 98.25% was reported. To prove the efficacy of REECGNet the authors added noise to the signals, and they obtained a recognition rate of 97.5%. Byeon *et al*. [1] used the same databases as in [2] to develop an identification system based on scalograms. The authors used three neural networks (AlexNet, GoogLeNet and ResNet) for comparison of the performances between the two databases. The highest accuracies were obtained with ResNet for both databases. Accuracies of 98.10% and 93.20% were reported for PTB-ECG and CU-ECG, respectively.

3 Dataset and Methods

In this section, the dataset and methods used in this work are described.

3.1 Dataset, Signal Processing and Biometric Models

The data used in this work was obtained from the database used in [5]. The ECG signals from 20 participants (11 females and 9 males aged between 19 and 28 years old), in neutral condition, were used. ECG data has approximately 5 minutes long, and two records per person, equally spaced in time, were used: one for training and the other for the test set. All individuals were healthy, and the ECG records were obtained following the Lead II configuration with a sampling frequency of 1000 Hz.

To filter the signals we used a sixth-order pass-band Butterworth filter with cutoff frequencies of 0.5 Hz and 30 Hz. After filtering the signals,the cardiac cycles were extracted according to the Hamilton's method [1]. Based on the R peak in each heartbeat, it extracts the cardiac cycle considering 0.2 seconds before and 0.4 seconds after the peak. Then the 10 and 20 more similar cardiac cycles for each person in both training and test sets were obtained. Afterwards, the scalogram of each of those cycles was computed. A scalogram is the absolute value of the continuous wavelet transform coefficients of a signal [1], allowing to obtain information in both time and frequency domains. In Figure 1 is presented a cardiac cycle of an individual and its respective scalogram.

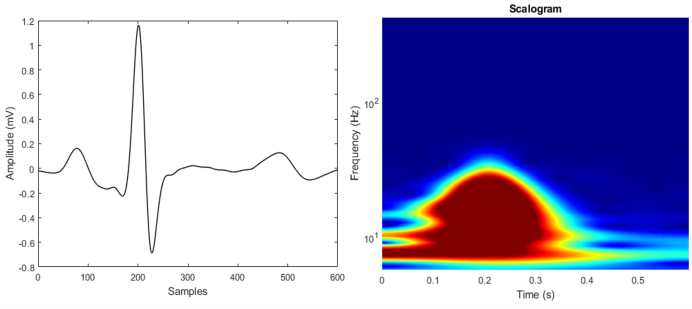


Figure 1: Cardiac cycle and respective scalogram

[1] Available through the neurokit package for Python

3.2 Identification and Authentication Tasks

The cardiac cycles and scalograms obtained were used as biometric models for both identification and authentication tasks, allowing a comparison between the two models in both tasks.

We calculate the distances between the models in the test and the training sets using two distances: the Manhattan distance and the Euclidean distance.

For the identification task, the classification was performed by the 1-nearest neighbour. For each test model of an individual, we computed the minimum distance to all models contained in the training set. So, for each test model, it was determined the individual to whom the closest training model belonged. For each type of model, the identification of the person was determined by the mode of the identifications obtained.

Regarding authentication, we computed a threshold for each participant. The threshold was obtained by the difference of the average distance between all the test models of each individual and all the training models of all individuals and its standard deviation. Then, all distances between the test and the training models of each individual were compared to their threshold. If more than 50% of these distances were lesser than the obtained threshold, the individual was authenticated. The Leave One Out method was used to evaluate the imposter rate. According to this evaluation strategy, each individual was removed from the training set, being the test set formed only by the individual removed. Authentication proceeded in the same way as previously described. However, in this case, if the individual was authenticated it was considered an imposter since it was no longer in the training set. This process was applied to all individuals.

4 Results and Discussion

This section presents the results obtained using the two models described.

4.1 Results using cardiac cycles and scalograms

Using cardiac cycles, we obtained identification and authentication rates of 100%, regardless of the distance or the number of cycles used. All the 20 individuals were correctly identified and authenticated. The highest imposter rate (14.7%), using these models, was obtained using Manhattan distance and 20 cycles. It decreased to 14.2% when using the Euclidean distance and 20 cycles or when we use the Manhattan distance and 10 cycles. The lowest imposter rate (13.7%) was obtained using the Euclidean distance and 10 cycles.

Using scalograms the highest identification rate (90%) was obtained when comparing 10 scalograms through the Manhattan distance. It decreased to 85% when we used Euclidean distance and 10 scalograms or when using Manhattan distance and 20 scalograms. The lowest identification rate (75%) was obtained comparing 20 scalograms using Euclidean distance. Concerning authentication, the highest rate (100%) was obtained using Euclidean distance, regardless of the number of scalograms used. It decreased to 95% when we use the Manhattan distance. However, using the Euclidean distance we obtained higher imposter rates than when we used the Manhattan distance. When using Euclidean distance, we obtained an imposter rate of 15.5%, regardless of the number of scalograms. On the other hand, when we used the Manhattan distance the imposter rate equalled 12.6% and 12.4%, with 10 and 20 scalograms, respectively. Figure 2 summarizes the results described.

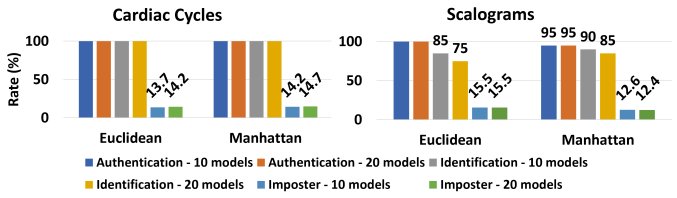


Figure 2: Results obtained using cardiac cycles and scalograms

4.2 Discussion

The results presented above indicate that the best distance and number of models to use depends on the biometric task intended to perform.

Despite being a model with more information than cardiac cycles, the results obtained with scalograms are lower than those obtained with cardiac cycles.

In general, when comparing cardiac cycles using the Euclidean distance, we achieved the lowest imposter rates. On the other way, using scalograms as biometric models, the Manhattan distance provides a lower imposter rate.

Using cardiac cycles, the best results were obtained comparing 10 cycles through the Euclidean distance, obtaining maximal identification and authentication rates (100%), with an imposter rate of 13.7%. Using scalograms, the best identification rate (90%) is achieved comparing 10 scalograms using Manhattan distance. Regarding the authentication task, we need to establish a trade-off between authentication and imposter rates. Therefore, we consider to slightly compromise the authentication rate to obtain a lower imposter rate comparing 20 scalograms through the Manhattan distance. This way, we achieved an authentication rate of 95% and an imposter rate of 12.4%.

5 Conclusion

In this work, biometric identification and authentication tasks, using ECG, were successfully applied. Regarding the identification task, the identification rate could be improved by using another classifier or, at least, using more nearest neighbours.

Although restricted to the size of the dataset and to the location of the acquisition of the signals, the obtained results allow to highlight the potential of the proposed biometric tasks. We used signals recorded at the wrist which are less susceptible to noise than those recorded at the fingers. On the other hand, the acquisition at the fingers is minimally intrusive being a better option for biometric systems. Therefore, we consider to create a dataset designed for biometric purposes, collecting ECG data in fingers or hand-palms. Despite the local of acquisition and the size of the dataset is different, our method seems to lead to better results than those obtained in the works [3] and [2]. Comparing to the article [1] we obtained lower identification rates than the authors.

Nevertheless, the high identification and authentication rates obtained are encouraging enough to further perform a comparison study with similar works comparing the methods used and applying them to different databases.

6 Acknowledgment

This work was funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the Scientific Employment Stimulus - Individual Call - CEECIND/03986/2018, and by the FCT through national funds, within IEETA/UA R&D unit (UIDB/00127/2020).

References

[1] Yeong Hyeon Byeon, Sung Bum Pan, and Keun Chang Kwak. Intelligent deep models based on scalograms of electrocardiogram signals for biometrics. *Sensors*, 19(4), 2019. ISSN 14248220.

[2] Jae Neung Lee and Keun Chang Kwak. Personal Identification Using a Robust Eigen ECG Network Based on Time-Frequency Representations of ECG Signals. *IEEE Access*, 7:48392–48404, 2019. ISSN 21693536.

[3] André Lourenço, Hugo Silva, and Ana Fred. Unveiling the biometric potential of finger-based ECG signals. *Computational Intelligence and Neuroscience*, 2011, 2011. ISSN 16875265.

[4] Mariusz Pelc, Yuriy Khoma, and Volodymyr Khoma. ECG signal as robust and reliable biometric marker: Datasets and algorithms comparison. *Sensors*, 19(10), 2019. ISSN 14248220.

[5] Gisela Pinto, João M. Carvalho, Filipa Barros, Sandra C. Soares, Armando J. Pinho, and Susana Brás. Multimodal emotion evaluation: A physiological model for cost-effective emotion classification. *Sensors*, 20(12):1–13, 2020. ISSN 14248220.

[6] Ranjeet Srivastva, Ashutosh Singh, and Yogendra Narain. PlexNet: A fast and robust ECG biometric system for human recognition. *Information Sciences*, 558:208–228, 2021. ISSN 0020-0255.

Low-Cost Pulse Oximetry and Infra-Red Temperature Device for COVID-19 Patients

Afonso Raposo[1,2]
afonso.raposo@tecnico.ulisboa.pt
Francisco Melo[2]
francisco.de.melo@tecnico.ulisboa.pt
J. Miguel Sanches[1]
jmrs@tecnico.ulisboa.pt
Hugo Plácido da Silva[2]
hsilva@lx.it.pt

[1] Institute for Systems and Robotics (ISR), LARSyS, Instituto Superior Técnico, Departamento de Bioengenharia, Universidade de Lisboa
[2] IT - Instituto de Telecomunicações
Instituto Superior Técnico
Lisbon, PT

Abstract

With the beginning of the COVID-19 pandemic in early 2020, there was a pressing need for simple yet effective remote monitoring solutions. In this paper, we describe a low-cost device developed for monitoring COVID-19 patients. The device uses an ESP32 module and integrates two distinct off-the-shelf biomedical sensors: a pulse oximeter by MAXIM, and an infra-red (IR) thermometer by MELEXIS. The device communicates with a smartphone via Bluetooth which then sends the acquired data to a cloud-based platform. An initial evaluation was performed at Coimbra's Polytechnic Institute, and covered reproducibility and agreement with standard clinical devices, revealing a strong correlation for the pulse oximeter and a necessity for further testing the IR thermometer.

1 Introduction

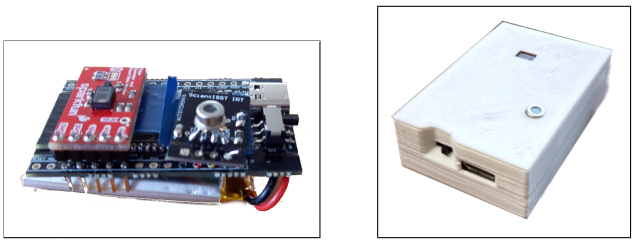
The last two years were marked by the COVID-19 pandemic that struck the world. This disease, caused by the virus SARS-CoV-2, spread worldwide and forced health care systems to quickly adapt to a remote monitoring paradigm. Lacking the tools to do so, the majority of the monitoring was performed through phone calls between the patients and the health care providers, further increasing the overall workload of the latter.

Phone calls were used to query patients regarding their symptoms and vital signs, trying to grasp the evolution of the disease. Patients would be questioned about the most frequent COVID-19 symptoms, namely: tiredness, dry cough, myalgia, dyspnea, loss of smell, gustatory dysfunction, rhinorrhea, asthenia, and sore throat [1, 2, 3]. They would also be asked to evaluate their body temperature for tracking fever, which is a usual predictor for this disease. Knowing the blood oxygen saturation would allow detecting "silent" hypoxemia (*i.e.*, unperceived lack of oxygen) [5], something that can occur in COVID-19 patients.

We developed a novel solution for this remote monitoring paradigm, which we called e-CoVig. Our solution implements at-home support and self-reporting of patients, using the smartphone as the primary data collection interface, coupled with a low-cost specialized device. We focused on the monitoring of heart rate (HR), oxygen saturation (SpO_2), body temperature, and symptomatology questionnaires [4]. Therefore, the same assessment performed through phone calls is still being carried out but, using the tools we developed, it is facilitated, more objective, and automated.

2 e-CoVig Device Implementation

The e-CoVig device, showcased in Figure 1, uses the ESP32 microcontroller as its mainboard and the off-the-shelf sensors connected to it. A pulse oximeter, the MAX30101 (high-sensitivity SpO2 and HR sensor using reflective photoplethysmography), is coupled with the MAX32664 (low-power sensor hub family, which seamlessly communicates with several MAXIM biometric sensors and computes biometric information). An IR thermometer, the MLX90615, which we integrated into our own PCB, is used for body temperature measurement. The temperature measurements are smoothed using a moving average filter (N=3), and the body temperature from the measured surface temperature is estimated using linear regression. The parameters for this regression were obtained by comparing temperature measurements between the MLX90615 sensor and an F102 forehead IR thermometer.



(a) (b)

Figure 1: The developed device with pulse oximeter and IR thermometer (a), and the same device inside a 3D printed case (b).

Type of measurement	1st measurement	2nd measurement	Mean difference	p-value
Heart rate (BPM)	80.10 ± 15.60	81.00 ± 16.00	0.90 ± 4.30	0.257
SpO_2 (%)	97.30 ± 2.60	97.40 ± 2.10	0.10 ± 1.30	0.783
Temporal temperature (°C)	36.68 ± 0.26	36.75 ± 0.25	0.07 ± 0.16	0.020
Tympanic temperature (°C)	35.59 ± 0.24	35.56 ± 0.23	-0.03 ± 0.08	0.045

Table 1: Pairwise comparison of the e-CoVig repeated measurements.

The device measures SpO2, HR, and temperature two times per second and, once a Bluetooth connection is established, the values are sent via Bluetooth Serial in a JSON format to the smartphone, as shown below:

```
{"HR": 72, "Confidence": 100, "SpO2": 99, "Status": 3,
"Object Temperature": 33.0, "Ambient Temperature": 15.1}
```

When the smartphone receives the measured data, forwards it to a remote monitoring cloud platform designed for health care professionals. This allows for an easy, scalable, and remote monitoring of a large number of patients.

3 Results

A cross-sectional study was designed to ascertain the reproducibility and accuracy of the measurements of heart rate, SpO_2, and body temperature (temporal and tympanic) with the developed device. A population of 30 volunteers was recruited among the students and staff of Coimbra's Polytechnic Institute, which we greatly acknowledge. The average age of this group was 23.33 ± 9.67 with 22 participants out of the 30 being female. None of the participants presented any COVID-19 symptoms during the study. All the evaluations were performed in the morning, in a laboratory with appropriate and controlled conditions. Every participant was seated comfortably for 10 minutes, to ensure the best measurement conditions. For the reproducibility assessment, repeated-measurements comparisons were performed for each participant. For the agreement assessment, one measurement was performed with the developed device and, right after, another with a standard clinically validated device.

3.1 Reproducibility Results

The e-CoVig device measurements assessing reproducibility are summarized in Table 1 and are showed in Figure 2. As demonstrated, no significant differences were observed for heart rate, oxygen saturation, and tem-

This work was supported by Fundação para a Ciência e Tecnologia (FCT) under the projects' reference: 255_596880547 e-CoVig; UIDP/50009/2020; UID/EEA/50009/2019; DSAIPA/AI/0122/2020 (AIMHealth) through IT - Instituto de Telecomunicações; and through LARSyS - FCT Plurianual funding 2020-2023; which is gratefully acknowledged.

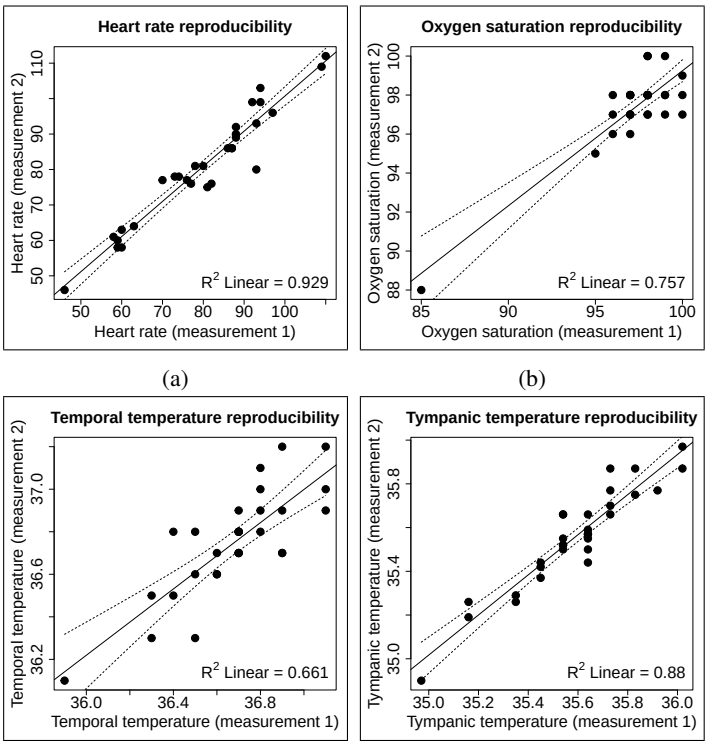


Figure 2: Regression plots representing the e-CoVig repeated-measurements correlation. (a) heart rate (in BPM); (b) SpO_2 (in %); (c) temporal and (d) tympanic temperatures (in °C). The solid line represents the regression line computed using unweighted least squares, while the dashed lines correspond to 95% confidence bands.

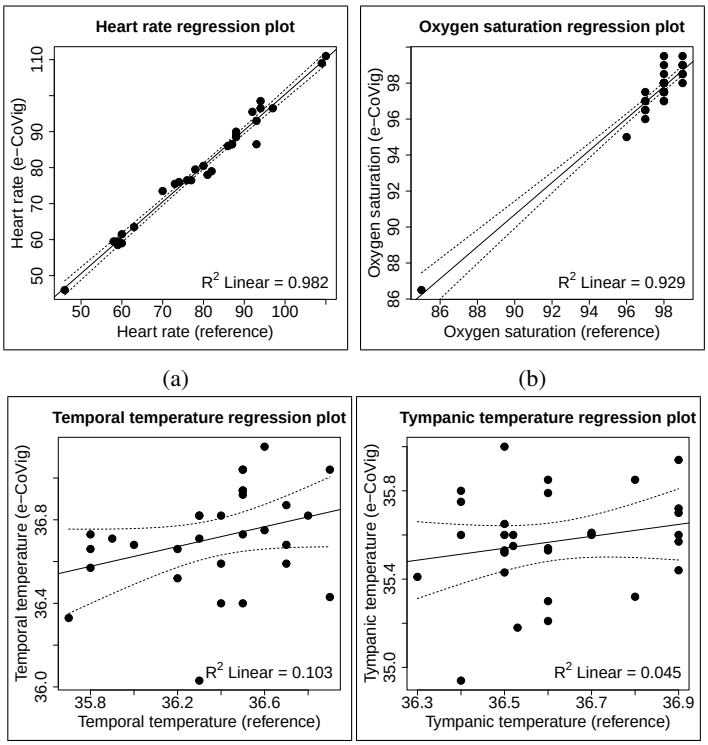


Figure 3: Regression plots representing the correlation between the measurements with the e-CoVig device and the standard reference device. (a) heart rate (in BPM); (b) SpO_2 (in %); (c) temporal and (d) tympanic temperatures (in °C). The solid line represents the regression line computed using unweighted least squares, while the dashed lines correspond to 95% confidence bands.

Type of measurement	Standard device	e-CoVig device	Mean difference	p-value
Heart rate (BPM)	80.10 ± 15.60	80.50 ± 15.70	0.40 ± 2.10	0.912
SpO_2 (%)	97.50 ± 2.50	97.40 ± 2.30	-0.10 ± 0.70	0.829
Temporal temperature (°C)	36.37 ± 0.33	36.71 ± 0.24	0.34 ± 0.34	< 0.001
Tympanic temperature (°C)	36.62 ± 0.18	35.57 ± 0.23	-1.04 ± 0.26	< 0.001

Table 2: Pairwise comparison between the e-CoVig and the standard device measurements.

poral and tympanic temperatures. It is important to note that the temporal temperature measurements may appear sparser, but the range of measurements is 1 °C and the IR thermometer used has an accuracy of ±0.5 °C.

3.2 Agreement Results

Regarding the comparison of measurements between the e-CoVig device and clinically-validated devices, the results are summarized in Table 1 and are presented in Figure 3.

The mean difference obtained for the e-CoVig tympanic temperature measurement is greater than the temporal temperature (-1.04 ± 0.26 vs. 0.34 ± 0.34, respectively, $p < 0.001$), which can be explained by the fact that the temporal measurements correspond to an estimation of the core temperature; in reality, the forehead is colder than the tympanic region. Therefore, since a linear transformation is applied by the e-CoVig device to estimate the body temperature, the device will overshoot for tympanic measurements. Nevertheless, the standard deviation was smaller for the tympanic measurements, indicating that this method of acquisition is less susceptible to variability caused by the acquisition procedure.

The temperature measurements presented worse results, which can be explained by the narrow range of temperatures measured, as explained in Section 3.1.

4 Discussion

Preliminary validation results with healthy patients indicate that it accurately and reliably provides measurements of heart rate, blood oxygenation, and body temperature. This device is easy to use and low-cost to produce, facilitating its integration in the remote monitoring paradigm of COVID-19 patients. Besides being a useful tool in households for monitoring people, it can also be used in nursing homes and other health care facilities, since it merges two common devices into one: a pulse oximeter and an IR thermometer.

The temperature measurements presented a narrow range of values in the trial at Coimbra's Polytechnic Institute, reinforcing the necessity of validating the e-CoVig device with a broader range of temperatures, allowing to validate this use-case for the MLX90615 IR temperature sensor. Fortunately, our device is currently being used at Santa Maria's Hospital with COVID-19 patients, which will allow for further validating the developed low-cost specialized device. Evaluating the preliminary results of 20 samples, we report an improvement in the temporal temperature measurements, showing a mean difference of -0.31 ± 0.50 and a p-value of 0.101.

References

[1] Wei-jie Guan et al. Clinical characteristics of coronavirus disease 2019 in china. *New England Journal of Medicine*, (18):1708–1720, 02 2020.

[2] Jerome R. Lechien. Clinical and epidemiological characteristics of 1420 european patients with mild-to-moderate coronavirus disease 2019. *Journal of Internal Medicine*, 288(3), 2020.

[3] James B. O'Keefe et al. Predictors of disease duration and symptom course of outpatients with acute covid-19: a retrospective cohort study. *medRxiv*, 2020.

[4] Afonso Raposo et al. e-covig: A novel mhealth system for remote monitoring of symptoms in covid-19. *Sensors*, 21(10), 2021.

[5] Ingrid Torjesen. Covid-19: Patients to use pulse oximetry at home to spot deterioration. *BMJ*, 2020.

Improving Federated Learning Protection with Digital Envelopes

Mario Dib[1]
mariodib@outlook.com
Pedro Prates[23]
prates@ua.pt
Bernardete Ribeiro[4]
bribeiro@dei.uc.pt

[1] Centre for Informatics and Systems of the University of Coimbra (CISUC), Coimbra, Portugal
[2] Department of Mechanical Engineering (CEMMPRE), University of Coimbra, Coimbra, Portugal
[3] Department of Mechanical Engineering, Centre for Mechanical Technology and Automation (TEMA), University of Aveiro, Aveiro, Portugal
[4] Centre for Informatics and Systems of the University of Coimbra (CISUC), Department of Informatics Engineering, Coimbra, Portugal

Abstract

The Federated Learning method was developed to to provide an alternative for the recent concerns with data privacy in machine learning. This method involves multiple parties to privately train local machine learning models with their own data, sharing with the global server only the models' parameters that will be averaged to update the global model. Although private, such environments are constantly at the risk of suffering cyber-attacks that can compromise the information used in the process and/or the complete machine learning training. This work investigates the application of Digital Envelopes combined with Federated Learning, to improve protection against attacks to either the clients and the server.

1 Introduction

In the field of machine learning, Federated Learning (FL) brings a new concept that includes a privacy-preserving approach regarding the datasets used in the model's training, allowing multiple participants to collaborate with their data, without giving up their private information. This allows to solve common issues that would be more challenging if done alone, opening a new way to approach issues' solutions in the industrial field [4]. However, there are some concerns regarding data protection, since the FL approach is susceptible to cyber-attacks, that includes backdoor attacks [1], model poisoning [2], untargeted attacks [3] and data poisoning [6], which is the one analyzed in this work. These attacks can compromise the models' results. So, this work proposed an approach to handle the data poisoning attack before the machine learning training, in order to prevent the models to be compromised, by combining the federated learning approach with the digital envelopes (DE). That way, its possible to verify the authenticity of the client trying to participate in the training phase and the integrity of the datasets, rather than looking for malicious patterns in the data. The rest of the paper is organized as follows. Section 2 introduces background and Section 3 describes the proposed approach. Section 4 details the experimental setup. Section 5 presents an analysis on the experimental results and, finally, Section 6 concludes and delineates future research lines.

2 Background

2.1 Federated Learning

The FL method consists in having a central server that orchestrates the activities to be done. The machine learning model is created by the central server and sent to all participants (clients), that perform the local training only with their local datasets, assuring the data privacy, since the local data is kept hidden from the other participants. The trained parameters (models' weights) are sent back to the central server, where all parameters will be averaged by the federated averaging algorithm [4]. Then, the central server updates the central model and sends the model back to the participants until all the communication rounds are done [4]. Figure 1 shows the simplified FL architecture.

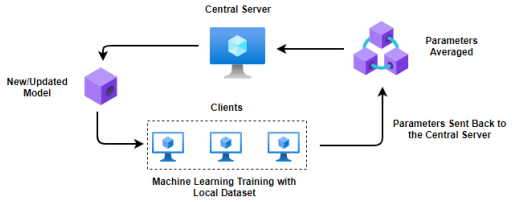


Figure 1: Federated Learning Architecture

2.2 Digital Envelopes

DE use asymmetric cryptography to protect the information. A random generated key, called secret key or symmetric, is used by the sender to encrypt an information. This key is also used for decryption by the receiver. After the information's encryption, the symmetric key is also encrypted but by the receiver's public key. Then, the encrypted document and the encrypted symmetric key are packaged together in a called digital envelope and sent to the receiver, and only the owner of the corresponding receiver's private key has access to the content of the digital envelope [5], as shown in Figure 2.

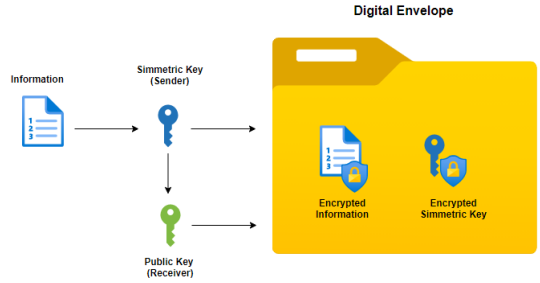


Figure 2: Digital Envelope Approach

2.3 Data Poisoning

Data poisoning attacks consists in maliciously manipulate the datasets in order to disturb the machine learning training. They can occur in two ways: causative attacks, where the attacker alters the training process through influence over the training data; and exploratory attacks, where the attacker does not alter the training process change misclassifications to compromise the model's performance [6], for example, the change of correctly labeled output with incorrect labels in order to disrupt performance and preventing targeting the desired objective.

3 Proposed Approach

In the proposed approach, there is a central server and clients (in this case 9 clients) to participate in the training procedure. The server holds its own private key and also all the clients' public keys, while the clients hold their own private key, the central server's public key and also generate the symmetric key. Each client encrypts its dataset with the symmetric key and then encrypts the symmetric key with the server's public key. In the next new step that is not part of the original digital envelope approach,

the hash of the encrypted document is created, thus adding a new layer of protection. The server then sends the created model to the clients and before the local training starts verifies two procedures inside the clients: whether the digital envelope belongs to its intended client and if it is the one created by the intended client, by checking if the envelope's hash corresponds to the client's keys. If negative, the clients are excluded from the training, otherwise, the digital envelope is then opened locally, and the standard FL approach continues. To allow the server to have a test dataset available for evaluating the models, a piece of each of the clients' dataset, as DE, is transferred to the server and only aggregated in run time memory, so the content is not accessible and its privacy is assured.

4 Experimental setup

Use Case. To test this approach, the selected use case was based on sheet metal forming processes, which are widely used in the automotive manufacturing sector to produce parts with complex geometries at high cadences. In this sector, process productivity is often impaired by sources of variability (e.g. material properties, process parameters), which may lead to forming defects and subsequent high scrap rates. Such challenges can be tackled in a collaborative manner with FL-based decision support tools, despite the strong data privacy policies in the sector.

Dataset. A dataset populated with numerical simulation results of the U-Channel forming process was used to evaluate and validate the proposed strategy. Details concerning the numerical modelling and simulation of this forming process can be found elsewhere. The features were related to elastoplastic behavior of materials and with the blankholder force. These numerical simulation consists in the geometric change in parts obtained by stamping, resulting from the elastic recovery of the material after removal of stamping tools, generating an output of 1 when the number is outside of a given allowable tolerance, meaning a defective part, or 0 when the part is considered normal. A well-known benchmark test to evaluate springback, namely U-Channel forming process, was used to generate the data. The dataset has 4200 samples divided in 3 material types: Mild Steel, DP600, and HSLA340, and was split for 9 clients, where each client received unique samples of one material type.

Experiment Specifications. To validate our approach, we have used the previously described dataset and a neural network with 5 layers for training the model, based on the horizontal federated learning architecture, which means that the different datasets need to have the same feature space. For the private and public keys generation, 2048 bit were used and 32 bit of length to randomly create the symmetric key, while other components, such as RSA, AES and ChaCha20 were used to complete the file encryption and sha3_512 to generate the hash.

Evaluation. In order to evaluate the binary decision task of the proposed models we defined well-known measures based on the possible outcomes of the classification, namely accuracy, which is the fraction of predictions the model classified correctly over the total number of predictions.

5 Experimental results and analysis

We evaluate the performance obtained on the data set by first creating the baseline model using the standard federated learning algorithm, where 3 models were trained, with different hyperparameters. The model with the best result was selected to be the baseline. To simulate the attacks, random clients were selected to have poisoned datasets and we analyzed the effect of these datasets during the training phase. These poisoned datasets are obtained by changing all the dataset's outputs to 0, meaning that the model would miss to classify springback. We observed that the more poisoned datasets participate in the training, the higher is their influence in the performance; therefore the accuracy decreases. However, when applied the digital envelopes, the malicious clients were identified correctly and excluded from the communication rounds achieving performances very similar to the baseline's performance. In which respects the effect of malicious samples in the test dataset, it also has the potential to obfuscate the results even when the models converged. Applying the digital envelopes also on servers allowed to obtain similar results comparing to the baseline, as can be seen in Table 1. The results showed that

Model	Malicious Datasets	Accuracy
Baseline	0	**0.9102**
Malicious - Train Dataset 1	1	0.8889
Malicious - Train Dataset 2	2	0.8534
Malicious - Train Dataset 3	3	0.7920
Malicious - Test Dataset 1	1	0.8582
Malicious - Test Dataset 2	2	0.8061
Malicious - Test Dataset 3	3	0.7589
FL+DE	3	**0.9113**

Table 1: Experimental Results.

data poisoning has the potential to cause damage by confusing the models with malicious data. During the training phase, when one malicious dataset was included in the training, a drop in accuracy from 0.9102 to 0.8889 can be seen, and the largest drop is when three malicious datasets were used, dropping the accuracy to 0.7920. When the models are trained correctly but the testing dataset is poisoned, leads to the understanding that the models are not performing well, with accuracy of 0.8582, 0.8061 and 0.7589 from lower to higher poisoned samples in the test dataset respectively. When the federated learning is used together with the digital envelopes, the achieved performance is similar to baseline model, with accuracy equal to 0.9113.

6 Conclusions and future work

We have proposed a method to improve the Federated Learning protection against data poisoning using Digital Envelopes. The combined approaches were able to identify correctly when a dataset did not belong to the expected client, removing it from the training and/or evaluation of the models, and therefore providing accuracy on par with the baseline model's accuracy. Also, no significant difference in performance was noticed, in terms of process time, when added the digital envelopes to federated learning framework, and since this approach is focused on identify the integrity and ownership of the datasets, rather than analyze the data itself, it would be possible to help identifying other types of attacks involving data. Our future work will include a more profound study about the application of this method against other types of cyber-attacks in order to evaluate its efficiency and also to extend the protection to the models used in the Federated Learning.

7 Acknowledgements

Research funded by FEDER and by FCT under the projects UIDB/00285/2020, UIDB/00326/2020, UIDB/00481/2020 and UIDP/00481/2020 and co-funded by POCI under the projects PTDC/EME-EME/31243/2017 (RDFORMING) and PTDC/EME-EME/31216/2017 (EZ-SHEET).

References

[1] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov. How to backdoor federated learning. *CoRR*, 2018.

[2] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo. Analyzing federated learning through an adversarial lens. *In International Conference on Machine Learning*, 2019.

[3] P. Kairouz, H.B McMahan, B. Avent, A. Bellet, M.Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cor-mode, and et.al. Advances and open problems in federated learning. *arXiv*, 2019.

[4] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas. Communication-efficient learning of deep networks fromdecentralized data. *arXiv*, 2016.

[5] S. Perez, J. L. Hernandez-Ramos, D. Pedone, D. Rotondiand, L. Straniero, and A. F. Skarmeta. A digital envelope approach usingattribute-based encryption for secure data exchange in iot scenarios. *Global Internet of Things Summit (GIoTS)*, pages 1–6, 2018.

[6] G. Sun, Y. Cong, J. Dong, Q. Wang, and J. Liu. Data poisoning attacks on federated machine learning. *arxiv*, 2020.

Road Accident Predictions as a Classification Problem

Madhulika Agrawal
magrawal@uevora.pt
Teresa Gonçalves
tcg@uevora.pt
Paulo Quaresma
pq@uevora.pt

Departmento de Informática,
Universidade de Évora, Portugal

Abstract

This paper aims at evaluating the performance of various classification methods for road accident prediction. The data is collected under MOPREVIS [3] project which aims at improving road safety in Portugal. The data is highly imbalanced as there are fewer accident instances than the non-accident ones and due to this imbalance, it is observed that the traditional classification algorithms do not perform well. Using sampling techniques (undersampling and oversampling) improved the results but not significantly. Some methods resulted in increased recall but that decreased precision as the algorithm returned more false positives to make up for data imbalance.

1 Introduction

According to the annual report published by Autoridade Nacional Segurança Rodoviária (ANSR) [1], in 2020 there were 26,501 accidents with victims on the continental Portugal, which resulted in 390 fatalities, either at the scene of the accident or during transport to the hospital. There were 1,829 serious injuries and 30,706 minor injuries.

In the period from January 1st to March 18th 2020, before the first period of lockdown resulting from the first State of Emergency, there was a general reduction in accidents when compared to the same period in 2019: 424 fewer accidents (-6.2%), 22 fewer fatalities (-22.0%), 41 fewer serious injuries (-9.6%), and 536 fewer minor injuries (-6.5%). In global terms, compared to 2019, in 2020 there was an improvement in the main accident indicators: 9,203 fewer accidents (-25.8%), 84 fewer deaths (-17.7%), 472 fewer serious injuries (-20.5%) and 12,496 fewer minor injuries (-28.9%).

Although there were fewer accidents in 2020, they are still not zero. And it is important for administration of the country to make transportation as safe as possible. In our research, our aim is to identify the time and spot that are most susceptible to accidents. This will help improving the road safety.

We discuss the problem definition in the next section which is followed by related work. Description of the dataset and the experimental setup, along with result discussion can be found in Section 3 and Section 4 respectively. The paper is concluded with scope for future work.

2 Problem Definition

There are several factors that could cause a road accident. Over the years, researchers are investigating the impact of these numerous variables and their significance in causing road accidents.

Road accident prediction can be defined as a two-class classification problem: an accident on a road, at any given time constitutes the positive sample; all the other times (when there are no accidents) are negative samples. The final problem definition is to identify if there will be an accident at a specific place and time instant.

By framing accident prediction as a binary classification problem, a major restriction arises: very high class imbalance. The number of instances when accidents happen are much lower than the non-accidents. The dataset built, described in the next section, has only 0.022% positive samples. This creates bias for the negative class while training a machine learning algorithm. One of the possible ways to take care of class imbalance is to create a balance between positive and negative samples by undersampling or oversampling [4] before training the machine learning model.

3 Related Work

There is plenty of research on how to manage data imbalance. The survey paper by Ander Carreño *et al.* [5] summarizes various class imbalance studies. There are precisely three types of methods to handle data imbalance [7]; data-level methods, algorithm-level methods and hybrid methods. The use of Random Forest (RF) classifier along with sampling techniques for imbalanced data has been explored by Chao Chan *et al.* [6]. They tested their model on 6 different dataset with varying degree of data imbalance.

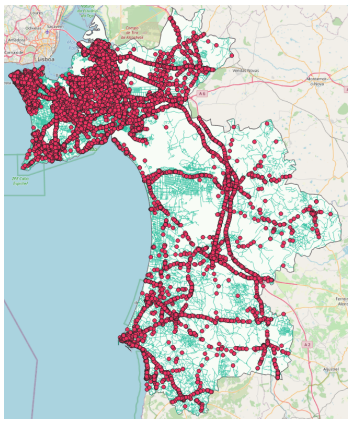


Figure 1: Distribution of Accidents in Setubal (PT)

Figure 2: Distribution of Accidents on N10 Road

4 Dataset

The accident database, provided by Guarda Nacional Republicana (GNR), consists of all the accidents that occurred in the Setubal region of Portugal between 1-January-2016 to 31-December-2019. The location coordinates of the accidents were mapped to the roads on Open Street Maps (OSM) [8]. For sake of simplicity, in our research we are focusing only on the accidents on National Road 10 (N10). Distribution of accidents in Setubal and N10 is presented in Figure 1 and Figure 2.

OSM divides the road into smaller, unequal segments. There are 186 road segments on N10 that had at least one accident in 4 years. The statistical analysis of number of accidents on various road segments and the length of those segments are presented in Table 1.

	#accidents	length (in meters)
mean	7.69	480.47
std	10.07	1027.19
min	1.00	21.21
25% percentil	2.00	82.29
50% percentil	4.00	186.87
75% percentil	9.00	464.29
max	71.00	8774.24

Table 1: Number of accidents per road segment and their length

For each road segment, the dataset built contains an entry for each hour of the day of the four years study. A sample is marked positive if an accident occurred in the one hour window for that day and is marked negative otherwise.

Along N10, Instituto Português do Mar e da Atmosfera (IPMA) have three weather stations. Depending on the proximity of the weather station to the road segment, the weather information for each instance was recorded from the nearest weather station. Other features, presented in Table 2, were also recorded and can be categorized as follows:

- **Time-Invariant Features:** Features of the road segment that do not change with time, like the presence of bridge or tunnel. There are 11 time invariant features;

- **Time-Variant Features:** Features that changes with time, like weather information. There are 12 time variant features associated with each segment of road.

	Feature	Description
Time Invariant	osm_id	unique ID of the road segment
	oneway	binary, if the road is a one-way
	bridge	binary, presence of bridge
	tunnel	binary, presence of tunnel
	codsubunidade	ID of the locality
	codpostal	postal code
	codconcelho	ID of concelho
	codfreguesia	ID of freguesia
	codsensepc	binary, if the road has central divider
	codtracado	binary, if the road is straight
	codtipoloc	binary, if the road is inside the city
Time Variant	ano	year
	ms	month
	di	day of the month
	hr	hour of the day
	day_of_week	sunday, monday, tuesday...
	daylight	dawn, morning, afternoon, dusk or night
	t_med	Average air temperature at 1.5m
	hr_med	average relative humidity
	dd_med	average wind direction
	ff_med	average wind intensity
	pr_dur_acc	accumulated precipitation duration
	pr_qtd_acc	accumulated precipitation quantity

Table 2: Features of the Dataset

There are features for which the values were missing for some hours; those data points we removed from the dataset. Table 3 details the size of the dataset, before and after removing data points with missing values.

	Unclean Data	Clean Data
Total Samples	6,521,904	6,402,059
Positive Samples	1427	1416
Negative Samples	6,520,477	6,400,643

Table 3: Description of the Dataset

5 Experimental Setup

Imbalance learn [2] is an open source library relying on Scikit Learn and provides tools for handling class imbalance in classification and includes implementation of various undersampling and oversampling methods. Before training a machine learning model, the samples are either under-sampled (from the majority class) or over-sampled (from the minority class) to create a balance between the two classes. We tested 10 undersampling, 5 oversampling and 2 joint under and over sampling methods on our data. The dataset is divided into train and test sets in the ratio of 70:30. All the methods are used with their default parameters. The performance of the algorithms are measured over AUC-ROC, Precision, Recall, and F1 Score.

	Algorithm	AUC	Prec	Rec	F1
RF	w/o Class Weights	.5190	**.1818**	.0381	.0631
	Balanced Class Weights	.6035	.0730	.2076	**.1080**
	Bootstrap Class Weights	.6035	.0726	.2076	.1076
Bal. RF	w/o Class Weights	.7364	.0005	.7780	.0011
	Balanced Class Weights	.7379	.0006	.7231	.0012
	Bootstrap Class Weights	**.7433**	.0005	**.7947**	.0011
	Easy Ensemble Classifier	.6939	.0004	.7589	.0008

Table 4: Results of Different Classifiers

The performance of different classification algorithms on the imbalanced dataset is presented in the Table 4. As can be seen, Balanced RF [6] outperformed all the other methods. This is why it is used as base classification algorithm for all the remainder of the experiments. The results of different sampling algorithms are presented in the table 5. Undersampling of the majority class gave better results than any other.

	Algorithm	AUC	Prec	Rec	F1
Oversampling	ADASYN	.5261	.0709	.0525	.0603
	BorderlineSMOTE	.5237	.0651	.0477	.0550
	RandomOverSampler	.6035	.0713	.2076	.1061
	SMOTE	.5249	.0677	.0501	.0576
	SVMSMOTE	.5702	**.0993**	.1408	**.1164**
Undersampling	AllKNN	**.7471**	.0005	.8019	.0011
	ClusterCentroids	.7289	.0005	.7780	.0010
	EditedNearestNeighbours	.7364	.0005	.7804	.0011
	InstanceHardnessThreshold	.7389	.0005	.7780	.0011
	NearMiss	.4937	.0002	**.9809**	.0004
	NeighbourhoodCleaningRule	.7438	.0005	.7923	.0011
	OneSidedSelection	.7361	.0005	.7708	.0011
	RandomUnderSampler	.7334	.0005	.7708	.0011
	RepeatedEditedNN	.7404	.0005	.7852	.0011
	TomekLinks	.7376	.0005	.7804	.0011
Joint	SMOTETomek	.5249	.0673	.0501	.0574
	SMOTEENN	.5321	.0868	.0644	.0739

Table 5: Results of Under and Over Sampling

Analysing the results, the following general observations can be made:

- **Low Precision:** Since there are more negative samples in the dataset hence there are higher chances of them being classified as false positives by the classifiers.
- **High Recall:** Fewer positive samples converts to fewer false negatives in classification and therefore higher recall.

6 Conclusion and Future Work

The higher recall by some methods is because the classifier is classifying more samples as positives which also results in increased false positives, causing reduced precision. From the results obtained, it is evident that portraying accident prediction as a classification task is not efficient. In future, it would be interesting to see what kind of results we could get by treating this as an anomaly detection problem, considering accidents as the abnormal behaviour in the regular road data.

7 Acknowledgment

This research is supported by the MOPREVIS - Modelação e Predição de Acidentes de Viação no Distrito de Setúbal, funded by Fundação para a Ciência e Tecnologia (FCT), reference FCT DSAIPA/DS/0090/2018 under the National Initiative on Digital Skills 2030, Portugal.

References

[1] Relatórios de sinistralidade. URL http://www.ansr.pt/Estatisticas/RelatoriosDeSinistralidade/Pages/default.aspx.

[2] Imbalance learn documentation. URL https://imbalanced-learn.org/stable/.

[3] Moprevis. URL https://moprevis.uevora.pt/.

[4] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced distributions. *ACM Computing Surveys (CSUR)*, 49(2):1–50, 2016.

[5] Ander Carreño, Iñaki Inza, and Jose A Lozano. Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. *Artificial Intelligence Review*, 53(5):3575–3594, 2020.

[6] Chao Chen, Andy Liaw, Leo Breiman, et al. Using random forest to learn imbalanced data. *University of California, Berkeley*, 110(1-12): 24, 2004.

[7] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[8] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

Contour Estimation and Delineation using Adaptive Periodic Cubic Splines

Paulo Salgado
https://www.citab.utad.pt/researcher/paulo-alexandre-cardoso-salgado
Engineering Department,
University of Trás-os-Montes e Alto Douro

Pedro Couto
http://www.citab.utad.pt/researcher/pedro-alexandre-mogadouro-do-couto
Engineering Department,
University of Trás-os-Montes e Alto Douro

Abstract

In this work a periodic cubic spline is used for contour estimation, delineation and representation. A computationally efficient method to dynamically adjust the spline nodes to the desired contour is proposed. Moreover, all the parameters are considered unknown and doesn't require any action from user. This adjustment of the spline to the desired contour is made using the gradient of the image pixels but, any other image property could be used. The result is an unsupervised parametric contour estimation and delineation. As an example of application, experiments reported in this paper address the problem of foetus boundary estimation.

1 Introduction

In short, splines are polynomials smoothly connected being their joining points called knots [1][2][3]. For a periodic cubic spline, each one of its segments is a third-degree polynomial and the it´s a closed structure (the last segment connects with the first one).

Many image processing applications use polynomial splines that have proven to be a useful tool for many purposes namely in medical image processing. Among those applications, contour estimation and delineation are one of the most important and addressed problems in medical image analysis, for example, for organ and foetus boundary location [4].

Similar approaches are Snakes [5] [6] [7] and deformable models [8] which have also been used in contour estimation [9] [10]. The drawback of these approaches is that they are non-adaptive since some, or all, of the parameters must be set *a priori* [4].

Splines are smooth and well-behaved functions (polynomials) that are continuously differentiable to (n-1) when dealing with splines of degree n [1]. Thus, cubic splines are more suitable for curve fitting due to the ease of data-interpolation, differentiation and integration, as they give a smooth response [11]. Splines are known for having excellent approximation and convergence properties making this parametric representation well suited to encode boundaries optimally.

In this work, an unsupervised method for contour estimation and delineation, where all the parameters are considered unknown, is proposed.

2 Proposed methodology

A periodic cubic spline is composed by n segments smoothly connected by knots and it is a closed structure. Each one of its segments is a third-degree (m) polynomial and, therefore, it means one would need 4 (m+1) coefficients to describe each segment.

To impose the continuity of the spline and its derivatives up to the second (m-1) order there is an additional smoothness constraint, so that, there will only be one degree of freedom per segment [1].

Let N be the set of n coordinate knots (x_k, y_k), $k = 1, \cdots, n$, which are part of a closed line L(r), in the space of continuous lines up to the 2nd derivative.

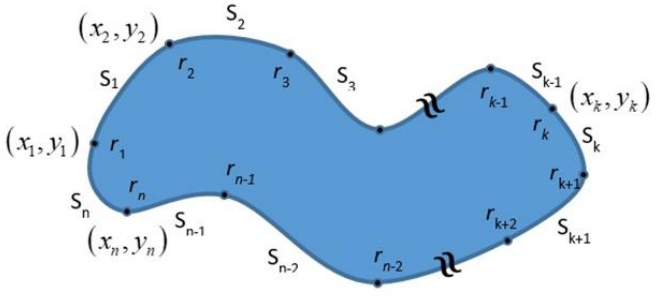


Figure 1: Spline contour representation

The line used is a periodic cubic spline (Figure 1). The cubic spline parametric equations are:

$$S_k(r) = \begin{cases} S_{x,k}(r) = a_{x,k}\Delta r_k^3 + b_{x,k}\Delta r_k^2 + c_{x,k}\Delta r_k + x_k \\ S_{y,k}(r) = a_{y,k}\Delta r_k^3 + b_{y,k}\Delta r_k^2 + c_{y,k}\Delta r_k + y_k \end{cases} \quad (1)$$

Being $\Delta r_k = r - r_k$ where r is an independent variable defined in the interval $r \in [0, r_{n+1}]$, for each r value there´s a corresponding point in the Line $L(r)$ with the coordinates $S_k(r) = \left(S_{x,k}(r) S_{y,k}(r)\right)$.

For increasing values of r, from 0 to r_{n+1}, the position of that point, $S_k(r)$, traverses the closed line of the initial node (x_1, y_1), passing through the intermediate knots, $(x_k, y_k)$for $r = r_k$, $k = 1, \cdots, n$, closing at the final knot, with the coordinates $(x_{n+1}, y_{n+1}) = (x_1, y_1)$, coincident with the initial knot, for $r = r_{n+1}$. However, the shape of the line L does not depend on the independent variable r.

The 1st and 2nd order derivatives of the component's Spline "x" are, respectively:

$$S'_{x,k}(r) = 3a_{x,k}\Delta r_k^2 + 2b_{x,k}\Delta r_k + c_{x,k} \quad (2)$$

and

$$S''_{x,k}(r) = 6a_{x,k}\Delta r_k + 2b_{x,k} \quad (3)$$

Spline component "y" has identical expressions.

The Spline curve obeys the condition that in its knots there will be continuity in the curve but also in the 1st and 2nd order derivative, ie:

$$S_k(r_{k+1}) = S_{k+1}(r_{k+1}) \wedge S'_k(r_{k+1}) = S'_{k+1}(r_{k+1}) \wedge S''_k(r_{k+1}) = S''_{k+1}(r_{k+1}), k = 2, \cdots, n \quad (4)$$

In the present case, the Spline is periodic (closed curve) and we also have that:

$$S_n(r_{n+1}) = S_n(r_1) \wedge S_n(r_1) = S_1(r_1) \wedge S'_n(r_1) = S'_1(r_1) \wedge S''_n(r_1) = S''_1(r_1) \quad (5)$$

For the n knots there are an equal number of Spline segments.

After determining coefficients, a, b and c [11], for the spline presented in (1), in this approach, a point on the section of the contour (ie, each Spline) varies its position between the extreme knots of the section, as the independent variable r changes from 0 to 1, $\alpha_k \in [0,1]$. So, for $r_k = 0$, that point it occupies the knot (x_k, y_k) and for $\alpha_k = 1$ that point occupies the next knot (x_{k+1}, y_{k+1}).

The regions of an image have different characteristics that can be described by power functions such as $G:(i,j) \mapsto \mathbb{R}_0^+$. For each pixel $(i,j) \in \mathbb{N}^2$ the power function provides a positive value that measures a feature or the aggregation of features of the pixel-centered region of the image.

The periodic spline (closed line) $L(r) = \bigcup_k S_k(r)$ crosses different regions of the image and inherits their characteristics. A method that moves the spline $S(r) = \left(S_x(r), S_y(r)\right)$ to regions with the desired image characteristics (gradient based in the exposed case) is proposed. The goal is, therefore, to maximize (or minimize) the value of the defined integral of the power function to which the Spline L line is subjected, that is:

$$\max_N J = \int_{L(r)} G\left(S_x(r), S_y(r)\right) dr \quad (6)$$

Being the (closed) spline $S = \left\{(x,y) : \left(S_{x,k}(r), S_{y,k}(r)\right), r \in [r_k, r_{k+1}] \wedge k = 1, \cdots, n\right\}$ with n sections and defined by the set of its non-knots. For a given set of knots, n, there is a spline line $S(x,y)$ that defines it. The shape of the line does not depend on the independent variable r, so this is not a variable to be optimized. The solution to the problem involves the optimal choice of the coordinates of the knots with $N = [\vec{x}, \vec{y}]$ that maximizes the function:

$$\max_N J = \sum_{k=1}^{n} \int_{r_k}^{r_{k+1}} G\left(S_{x,k}(r), S_{y,k}(r)\right) dr \quad (7)$$

This way, spline knot coordinates can be iteratively adjusted.

To accelerate the convergence of the method, a clustering algorithm is used. Let C be the matrix (image) with the values (pixels) of the highest power function value $C = max(0, G - \beta)$ with $\beta \in [0, max(G)]$.

Let n be the number of groups to be defined in C and N be the set of coordinates of the knots (in number equal to n) of the Spline. Let the active region be C (C>0). Spline knots behave like cluster centers in a given iteration. The "vector" of the knots will be the result of the adjustment of the area center against the distribution of the pixels (Algorithm K-means)

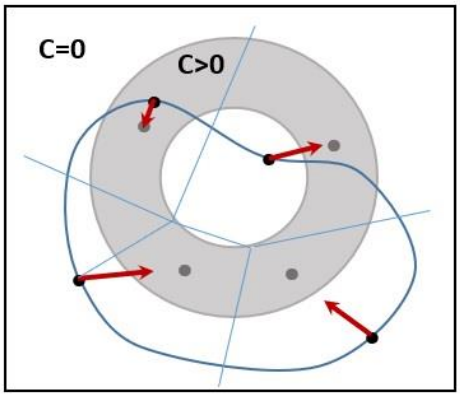


Figure 2: Spline contour synthetic example

In Figure 2 a synthetic representation of a contour is given where the • mark represents the knot new position after an iteration.

Let N be the positions of the Spline knots and, simultaneously, be the centers of the n clusters. Being $K = \{k : C(k) > 0\}$ the set of all data (pixels or points) such as C value is positive. The data is then divided over the n cluster centers and, with that purpose, the degree of belonging of each pixel k to the n clusters is determined, this being the value of the variable $u_{ik} = 1$ for $i = 1, \cdots, n$ and $k = 1, \cdots, m$ where m is the number of pixels such as C>0 and: $u_{ik} = 1$ if $d_{ik} < d_{jk}, \forall j \neq i$ for $i, j = 1, \cdots, n$ (being $u_{jk} = 0, \forall j \neq i$) where d_{ik} is the distance of data k to the centre of cluster i. Thus, the data k belongs to the ith grouping.

Finally, all cluster centres are updated and will become the new spline knots.

This way, the spline knots will iteratively adjust to the contour.

3 Results

The proposed algorithm, for the implementation of the previous described methodology, is iterative and continuously adjusts the spline according to a power function. The power function to which the Spline L line is subjected is image dependent and directly related with the desired image attribute. In this implementation the power function is gradient based.

Figure 3: Original and gradient foetal ultrasound image

The examples used in this article are foetal ultrasound images and the goal is to estimate the gestational sac boundary that holds the foetus (Figure 3).

Being the power function, a gradient based one, the image used to adjust the spline is a gradient image (Figure 4).

The knots initial position is random (Figure 5a), and they iteratively adjust to the contour to achieve the final contour delineation (Figure 5b) and thus its parametric representation.

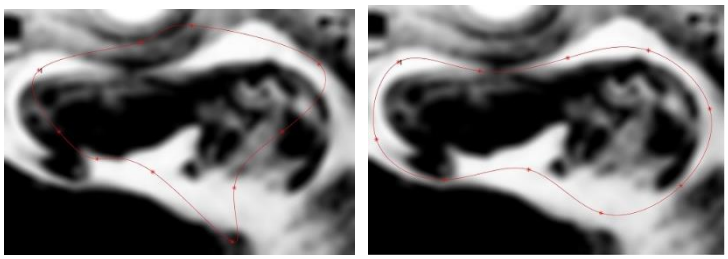


Figure 5: (a) spline initial knots random position and (b) spline final knots position and correspondent contour delineation

4 Conclusions

This is an unsupervised method where all the parameters are considered unknown and doesn't require any action from user except from the number of knots that are randomly positioned in the image.

Experiments prove that the proposed method can adaptively adjust the spline to the desired contour in an unsupervised way producing excellent results for dynamic boundary encoding.

Moreover, since the spline is adjusted according to a power function, this framework can be adapted to many other problems. For instance, autonomous robot trajectory estimation where the power function can be a safety trajectory grade.

Future work is intended to generalize the proposed framework to easily adapt it to other image and non-image related problems.

References

[1] M. Unser. Splines: a perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22-38, 1999. doi: 10.1109/79.799930.

[2] C. deBoor. A Practical Guide to Splines. New York: Springer-Verlag, 1978.

[3] P. Dierchx. Curve and Surface Fitting with Splines. Oxford, U. K. : Oxford University Press, 1993.

[4] M. A. T. Figueiredo and J. M. N. Leitão and A. K. Jain. Unsupervised Contour Representation and Estimation Using B-Splines and a Minimum Description Criterion. *IEEE Transactions on Image Processing.*, 9(6):1075–1087, 2000.

[5] K. Kass and A. Witkin and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1: 259–268, 1987.

[6] A. Samreen and M. Zawwar and I. Misbah. Image interpolation by rational ball cubic B-spline representation and genetic algorithm. Alexandria Engineering Journal. 57(2):931-937. 2018.

[7] S. A. Abdul Karim, "Rational Bi-Quartic Spline With Six Parameters for Surface Interpolation With Application in Image Enlargement," in IEEE Access, vol. 8, pp. 115621-115633, 2020.

[8] A. Yuille and P. Hallinan. Deformable Templates. *In Active Vision*, A. Blake and A. Yuille Eds., Cambridge, MA: MIT Press. 21–38, 1992.

[9] C. Xu and J. Prince. Snakes, shapes and gradient vector flow. *IEEE Transactions on Image Processing*. 7: 359-369. 1998.

[10] A. K. Jain and Y. Zhong and M. P. Dubuisson-Jolly. Deformable Template models: A review. *Signal Processing*, 71(1):109–129, 1998.

[11] C. Tutika and C. Vallapaneni and K. Ravichandran and K. P. Bharath and R. Nersisson and R. Muthu. Cubic Spline Interpolation Segmenting over Conventional Segmentation. *Procedures: Application and Advantages*. 2018.

[12] G. Micula and S. Micula. In: Handbook of Splines. Mathematics and Its Applications. 462: 383-600. Springer, Dordrecht. 1999. doi:10.1007/978-94-011-5338-6_12.

Regressing Autonomous Guided Vehicle Localization from Non-Visual Sensor Data

Bruno C. da Silva
http://www.di.ubi.pt/
Luís A. Alexandre
http://www.di.ubi.pt/~lfbaa/

NOVA LINCS and Departamento de Informática
Universidade da Beira Interior
6201-001 Covilhã, Portugal

Abstract

To navigate efficiently, a robot needs to have effective strategies regarding its navigation stages: perception, mapping, localization and path planning. In the localization aspect, a robot estimates its current location in an environment. The more precise this estimation is, the more accurate will be the map of the environment and the robot's ability to create a more precise trajectory of the path. In this paper we study different approaches to obtain an estimate of an autonomous guided vehicle localization, built from non-visual sensor data. We compare results from different regressions methods, namely ridge, lasso, elastic net and support vector regression, with data from individual sensors and two standard fusion approaches, Adaptive Monte Carlo Localization and the Extended Kalman Filter. We concluded that the elastic net regression is a viable method for fusion information from multiple sources (sensors and prediction algorithms) to improve the localization accuracy.

1 Introduction

Localization is a task which involves the robot to use its sensors to retrieve data from the environment and estimate its position. The measurements of its sensors are fundamental to help robots to perceive its surroundings and thus, perform the localization task.

Besides the level of sensor technology, the noise in these sensors must be taken into account to understand the difference between their measurement and the real world. Sensors such as inertial measurement units (IMU) and odometry can accumulate drift errors over time [2], and global positioning system (GPS) can suffer from signal propagation errors, dilution of precision and delays provided by earth layers [4]. Hence, the literature shows that the accuracy of the robot's position in an environment can be improved by integrating different sensor information.

This strategy has been adopted by several works throughout the years. Recently, Sarker et. al. [3] introduced a Bayesian filtering based data reconstruction scheme to increase the reliability of autonomous navigation of mobile robots. The authors transform the prediction step and propose an Imputation step of Extended and Cubature Kalman Filter models to work toward missing data estimation. In their evaluation, they compare the performance of the two Kalman-based methods using a baseline model data stream that uses only the unfiltered sensor data. They used localization coordinates calculated using data from a LIDAR, GPS module and orientation information provided by a gyroscope. Their method worked well for estimating unfiltered and corrupted data.

In the work of [1], the authors propose a lightweight algorithm which creates a virtual-IMU to store data from multiple IMU sensors. Their method fuses these information with exteroceptive sensors, achieving better localization accuracy compared to methods that fuses sensors with a single IMU and can be integrated with filter-based algorithms, as well as optimization-based filter algorithms. The simulated tests are performed using nine IMUs and a monocular camera, using poses and sensor measurement based on real-world data and the results show that the localization error can be improved. For the real-world tests, the authors recorded poses from a GPS-RTK module and used them as ground truth data creating three indoor and three outdoor datasets. Their tests regarding localization precision and improvement of the inertial odometry algorithm were successful.

Adaptive Monte Carlo Localization (AMCL) and Extended Kalman filter (EKF) are considered standard approaches in fusing sensors to achieve robot's localization. Although sensor-fusing techniques are popular for robot localization, distinct scenarios require the use of distinct approaches. In this paper we analyze different regression approaches, namely ridge, lasso, elastic net and support vector regression (SVR) to understand if these methods can be used to improve the robot's localization accuracy by fusing data from sensors compared with AMCL and EKF.

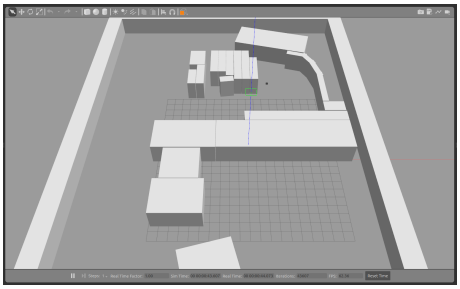


Figure 1: Simulated environment of a part of the Stellantis Factory in Mangualde, Portugal. This section represents the trajectory of where the AGV will navigate.

2 Our proposal

2.1 Simulated Environment

The evaluation was executed inside a simulated environment of the Stellantis factory in Mangualde, Portugal. The environment was created using Gazebo sim and ROS. Accordingly to the factory management, one way length of the trajectory is roughly 370m. With a speed around 0.2 m/s, it means that the AGV will take around 60 minutes do complete the path. The simulation was build in a scale of 10 times smaller.

An autonomous guided vehicle (AGV) inside a factory is used to transport loads without an onboard driver and its navigation aspects are software-sensor defined. For this work, the simulated autonomous capabilities of the AGV were configured using ROS packages such as the navigation package, which includes parameters to configure the move_base node and the Dynamic Window Approach algorithm to allow the robot to navigate autonomously. Since the data from the sensors and algorithms are in Cartesian format, the GPS latitude and longitude values were converted using ROS navsat_transform package.

The ground truth values were established by adopting the almost perfect odometry data from the Gazebo plugin. We compare values from localization-related sensors and fusion algorithms to the ground truth value and analyze the precision of each component. There are different sensors that can be used to measure data related to the robot position, the ones used for this work were divided in two parts: individual measurements and fusion-based measurements. For the individual measures, an odometry sensor and a GPS sensor were used. For the fusion-based measures, the AMCL algorithm was configured with odometry and laser sensors, and the EKF was configured with an IMU, odometry, and GPS sensors.

2.2 Data Extraction

The extraction of data from the sensors was done by creating a method to retrieve 30 Cartesian points per second from ROS topics. The Gazebo odometry, which has perfect location information, was used as our ground truth, an odometry configured with noise to simulate the output of real odometry sensors, a GPS sensor, the AMCL and EKF algorithms.

2.3 Fusing data with Regression Methods

After extracting data from the odometry with noise, GPS, AMCL and EKF we fused it using several regression methods: ridge, lasso, elastic net and support vector regression. The main goal is to investigate if these regression approaches could achieve more accurate results than the sensors and algorithms individually, therefore improving the robot localization.

First, we calculate the euclidean distance between the locations given by the ground truth and all the sensors and algorithms, and then we calculate the mean and the standard deviation of these results, which appear in the second columns of Table 1.

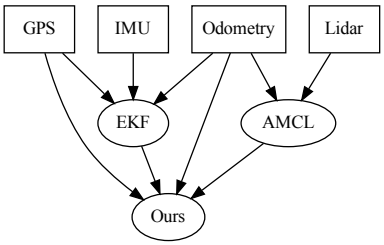


Figure 2: We propose to fuse localization information from two sensors (noisy odometry and GPS) and two prediction methods (AMCL and EKF). Rectangles represent sensors and ovals represent prediction methods. 'Ours' represents the regression approaches evaluated in this work.

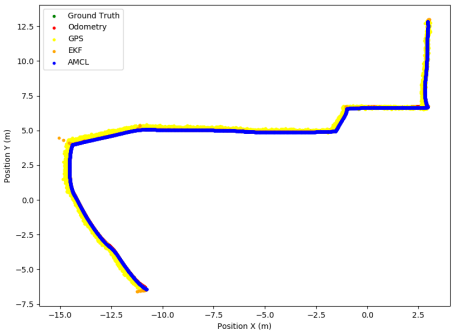


Figure 3: Comparison between the data from sensors/algorithms and ground truth values throughout the path navigated by the AGV.

With the extracted data, we created a data set containing 4989 localization points from the odometry (noisy and noise-free that is used as ground truth), GPS, AMCL and EKF. Figure 3 contains a visual comparison between the localization provided by the sensors, the algorithms and the ground truth values throughout the path in the simulated environment.

With these points, we performed tests with the regression models using the first 60% of the data for the training set, the following 10% for validation and remaining 30% for test. We did not shuffle the data such that no nearby points from one set appear in any of the others. As the regression models require a penalty term (C for SVR and λ for the others) to reduce bias and overcome overfitting, we needed to find the best value for these terms. We created a function where we tested a range of values for C and λ from the following set, $\{0.01, ..., 1\}$, to see which one would give us the best score for each regression model. This parameter optimization was performed using only the training and validation data. After that, we tested the models using the test data set containing the last 30% of the data extracted, with the following configurations: ridge, λ=0.2, *max_iter*=*None*, *tol*=0.001; lasso, λ=0.01, *max_iter*=1000000, *tol*=0.001; elastic net, λ=0.01, *max_iter*=10000, *to*l=0.001; and support vector regression, C=0.7, *max_iter*=1000.

The values for C and λ indicate the regularization penalty to improve the model estimate capacity by reducing its variance. The *max_iter* values are the number of iterations of the solver in the algorithm. Finally, the *tol* value is the precision of the solution, where a tolerance criteria for stoppage is established. The results of the optimization process are shown in Fig. 4. We can see that SVR is only affected by the C term for values below 0.07 and ridge regression is not affected by the particular value of λ on the tested range, while for the other two approaches, the λ that maximized the score was chosen. The values of *max_iter* were obtained in a similar manner starting from 1 and increasing tenfold until convergence was obtained. Finally, we used the default value for *tol*.

In Table 1, one can visualize the Root Mean Squared Error of the x (RMSE_x) and y (RMSE_y) coordinates and the Mean Absolute Error of x (MAE_x) and y (MAE_y) coordinates produced by the two sensors, the two estimation algorithms and the regression methods, on the test data.

The results show that fusing data from GPS, Odometry, EKF and AMCL using regression methods, can generate a more precise localization than when only using the original four localization sources separately. We also tested fusing only sensor data and only algorithm data to determine if the accuracy of fusing raw and transformed sets of data separately could surpass the proposed model. Despite the fusion between AMCL and EKF data presenting quite accurate localization values, when fusing all the sources, elastic net presented the best overall results, beating all other methods with smaller mean distance error and obtaining good results on the corresponding RMSE and MAE metrics w.r.t the X axis.

Table 1: Values for the mean and standard deviation of the euclidean distance between ground truth points and the sensors and algorithms points, root mean square error of x and y points, mean absolute error of x and y coordinates of the extracted data from the AGV's sensors and algorithms evaluated on the test data set.

	Mean (Stddev)	RMSE_x	RMSE_y	MAE_x	MAE_y
AMCL	0.235 (0.007)	0.094	0.216	0.092	0.216
Odometry	0.058 (0.006)	0.046	0.036	0.046	0.035
GPS	0.294 (0.051)	0.213	0.210	0.203	0.202
EKF	0.423 (0.290)	0.320	0.401	0.233	0.346
Ridge (all sources)	0.045 (0.008)	0.028	0.037	0.027	0.034
Ridge (GPS+Odom)	0.051 (0.008)	0.037	0.036	0.037	0.034
Ridge (AMCL+EKF)	0.076 (0.052)	0.022	0.089	**0.018**	0.071
Lasso (all sources)	0.052 (0.027)	**0.021**	0.054	**0.018**	0.045
Lasso (GPS+Odom)	0.066 (0.025)	0.061	0.035	0.055	0.033
Lasso (AMCL+EKF)	0.079 (0.051)	**0.021**	0.092	**0.018**	0.074
ElasticNet (all sources)	**0.043** (0.017)	**0.021**	0.041	**0.018**	0.037
ElasticNet (GPS+Odom)	0.065 (0.025)	0.062	0.032	0.055	0.029
ElasticNet (AMCL+EKF)	0.079 (0.051)	**0.021**	0.092	**0.018**	0.074
SVR (all sources)	0.047 (0.023)	0.045	**0.026**	0.040	**0.020**
SVR (GPS+Odom)	0.067 (0.023)	0.036	0.061	0.029	0.056
SVR (AMCL+EKF)	0.135 (0.024)	0.117	0.073	0.116	0.061

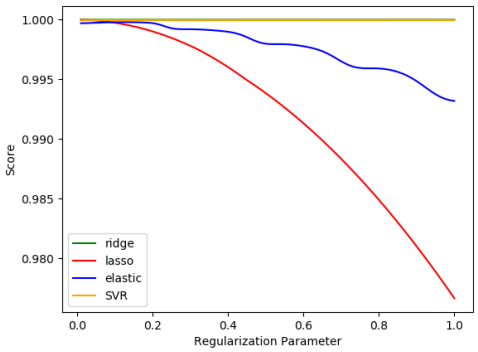


Figure 4: Scores while optimizing λ and C on validation data.

3 Conclusions

In this paper we explored the use of regression approaches to improve the localization accuracy of a simulated AGV. Our experiments showed that the elastic net regression method can be used as fusion method that can improve the localization quality of an AGV. This is somewhat expected as elastic net is useful for problems with multiple correlated features, which is the case here. Future work will explore the use of neural-based approaches to this problem.

Acknowledgments

This work was supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT- Fundação para a Ciência e a Tecnologia, through national funds, and partially supported by project 026653 (POCI-01- 0247-FEDER-026653) INDTECH 4.0 – New technologies for smart manufacturing, cofinanced by the Portugal 2020 Program (PT 2020), Compete 2020 Program and the European Union through the European Regional Development Fund (ERDF).

References

[1] Mingyang Li, Ming Zhang, Yiming Chen, and Xiangyu Xu. A lightweight and accurate localization algorithm using multiple inertial measurement units. *IEEE Robotics and Automation Letters*, 2020.

[2] Prabin Kumar Panigrahi and Sukant Kishoro Bisoy. Localization strategies for autonomous mobile robots: A review. *Journal of King Saud University - Computer and Information Sciences*, 2021.

[3] Victor Kathan Sarker, Prateeti Mukherjee, and Tomi Westerlund. Enhanced reliability of mobile robots with sensor data estimation at edge. *IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, 2020.

[4] Junjie Shen, Jun Won, Zeyuan Chen, and Qi Alfred Chen. Drift with devil: Security of multi-sensor fusion based localization in high-level autonomous driving under gps spoofing. *USENIX Symposium*, 2020.

Ultrasound denoising using the pix2pix GAN

Afonso Raposo[1]
afonso.raposo@tecnico.ulisboa.pt
António Azeitona[1]
antoniorrazeitona@tecnico.ulisboa.pt
Manya Afonso[2]
manya.afonso@wur.nl
J. Miguel Sanches[1]
jmrs@tecnico.ulisboa.pt

[1] Institute for Systems and Robotics (ISR), LARSyS, Instituto Superior Técnico, Departamento de Bioengenharia, Universidade de Lisboa
[2] Wageningen University and Research, Wageningen, The Netherlands

Abstract

The use of ultrasound (US) as an imaging technique is essential for the diagnosis of atherosclerotic cardiovascular disease (ASCVD), which depends on US images of the carotid artery. However, US images are plagued by a specific type of noise called Speckle noise, which lowers image quality dramatically. As an attempt to improve US image quality, the use of a generative adversarial network (GAN) is explored. The GAN chosen for this is the pix2pix model and the dataset used for training is composed of images containing simple geometric shapes of various scales and their equivalent corrupted with Speckle noise following the Log-Compression model. The results of this GAN are displayed and a noticeable improvement can be verified in the image quality.

1 Introduction

The two main predictors used for the diagnosis and assessment of atherosclerotic cardiovascular disease risk are the carotid intima-media thickness and analysis of the carotid arterial plaque, both of which are obtained by the use of ultrasound (US) imaging [4, 6]. Although there exist some promising studies on the development of a fully automatic segmentation technique, the performance is still far from ideal due to the high content of Speckle noise [5]. The current approach for the denoising of US images is based on the use of non-linear filters such as anisotropic diffusion filters and adaptive median filters [2]. These filters tend to preserve the contours of the structure but over-smooth the remaining areas.

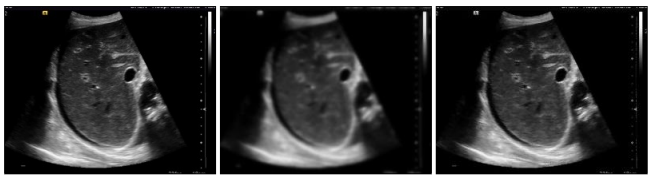


Figure 1: Ultrasound image of a liver (left) and corresponding images resulting from anisotropic diffusion filtering (middle) and adaptive median filtering (right).

The introduction of Generative Adversarial Networks (GANs) as a means to generate images presents a new opportunity for developing novel denoising techniques. The most widespread of these networks is the *pix2pix* [3], which can be trained with pixel-wise paired images in a way that it can receive a certain image as an input and then output a version of that image with different characteristics, image-to-image translation.

2 Problem Formulation

The pix2pix network requires pairs of images to be trained. In this case, these pairs consist of US images with Speckle noise and the same image without Speckle noise.

Speckle noise follows the Rayleigh distribution:

$$\rho(y_i) = \frac{y_i}{\sigma^2} e^{-\frac{y_i^2}{2\sigma^2}} \quad (1)$$

This work was supported by Portuguese funds through FCT (Fundação para a Ciência e Tecnologia) through the projects reference UIDP/50009/2020 and through the reference UID/EEA/50009/2019, LARSyS - FCT Plurianual funding 2020-2023.

Where ρ is the p.d.f., y_i is the intensity value of the ith pixel in the grayscale ultrasound image and σ is a scale factor dependent on the scattering amplitude of the particles in the medium [9].

B-mode US images suffer logarithmic compression after the acquisition of the data, which can be modeled by the equation 2.

$$z_{ij} = \alpha \log(y_{ij} + 1) + \beta \quad (2)$$

Where i and j are the positions of the pixel, z is the pixel after the compression, y is the pixel of the radio frequency (RF) image, and α and β are parameters dependent on the contrast and brightness [7], respectively.

These mathematical models make possible the creation of synthetic pairs of images to train the network, which, after trained, will accept US images and return denoised versions of those images.

3 Methods

The dataset used was composed of synthetic images of several geometric shapes of varying dimensions, intensities, and number. These images were generated with the *draw.random shapes* function from the *skimage* library, using the parameters: shape=(256, 256), allow_overlap=True, min_shapes=128, max_shapes=256, min_size=10, max_size=50. The intensity of the images was then inverted, so the background was black and the pixels with the lowest intensities (<5) were corrected to have an intensity of 5. A total of 2560 images was generated this way, corresponding to the output of the training dataset (denoised US images).

The training input images (noisy US images), were computed based on the images obtained using the method described above. To simulate the US image with Speckle noise, the Rayleigh distribution function (equation 1) was used, taking the original synthetic image as the value for the standard deviation (here denoted by σ). However, the model for logarithmic compression as displayed in figure 2, shows that after the noisy data is obtained (RF image), there are a few steps to reach the final B-mode US image.

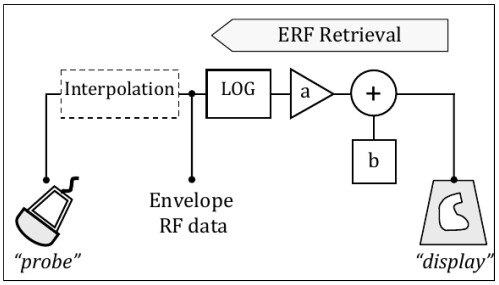


Figure 2: Full diagram of the model for the generic processing operations of an ultrasound system [8].

The first of these steps is an interpolation, which was mimicked by applying a 2D decimation of the images, reducing them to a quarter of their size, followed by applying a linear interpolation, restoring the dimension of the images.

The other step is the logarithmic compression expressed by equation 2. This process depends on two parameters, α and β, that are usually not provided by US equipment manufacturers and, therefore, are unknown. As a way to increase the versatility and robustness of the network, these parameters were randomly selected for each image from a set of intervals: $[10, 50]$ for α and $[-50, 50]$ for β.

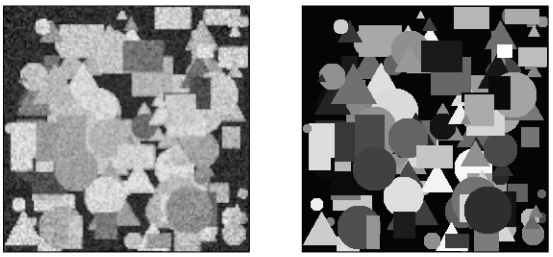


Figure 3: Example of the image pairs resulting from the method used: the input image (left) and the target (right).

The training of the network consisted of using 2048 image pairs for training and 512 image pairs for validation, for a total of 200 epochs.

4 Results and Discussion

After the network was trained, some images from the validation set were fed to it so that the image generated could be compared to the target output, as shown in figure 4.

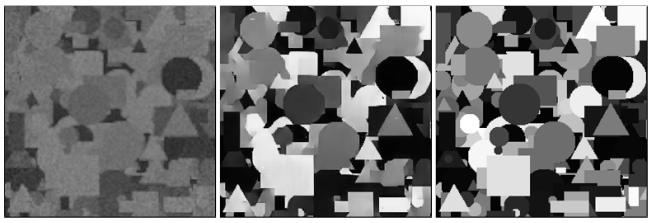


Figure 4: Example of a simulated noisy US image from the validation set (left), the output given by the trained network (middle) and the original synthetic image (right).

Although some distortion is present, the structural information recovered seems to be more than satisfactory and the values of the Structural Similarity Index (SSIM, higher is better) and Peak Signal-to-Noise Ratio (PSNR, higher is better) as quality assessment metrics were calculated [1] and the values are shown in table 1.

Table 1: Values of peak signal-to-noise ratio and structural similarity index corresponding to different denoising techniques

Method	PSNR	SSIM
Anisotropic Filter	11.906	0.461
Adaptive Median	12.053	0.452
Ours	21.085	0.789

The drastic difference in the values can be explained easily when comparing the images (figure 6), seeing as the classical filters do not improve the intensity values of the image in the same way that the network was able to.

Nonetheless, both the SSIM and PSNR, testify to the potential of the network when compared to classical methods of filtering, showing almost double the score.

The finished network was also used to denoise real US images of the carotid artery, resulting in the images shown in figure 5.

In this case, quality assessment measures cannot be performed because, being a real US image, there is no ground truth image (a completely clean image) available. Even so, visual comparison with the aforementioned classical methods is possible (figure 6).

As it can be seen, the image resulting from the GAN highlights the more prominent structures, preserves most contours without retaining noise while, admittedly, losing some of the "realness" of the image as a default US image since the model was trained with geometric synthetic images.

5 Conclusion

The use of the pix2pix network as a tool to denoise and enhance the image quality of US images is an appealing prospect and, as shown in this work, reveals itself promising in this area. In this work, synthetic images were

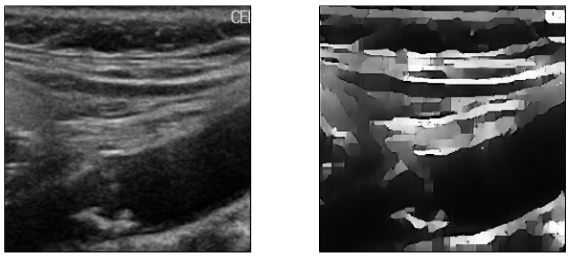


Figure 5: Example of a real carotid US image from (left) and the output given by the trained network (right).

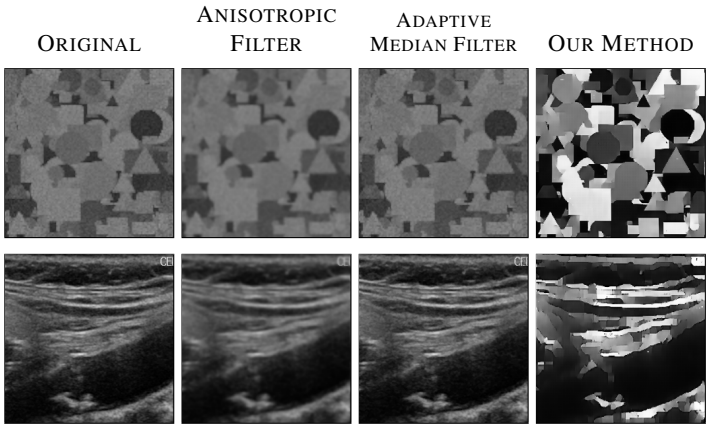


Figure 6: Comparing various ultrasound denoising algorithms to our method.

used to train the network and, although the influence of the geometric shapes is clear on the resulting carotid US image, the increased sharpness and preservation of contours holds great value for both physicians and automatic segmentation algorithms. There is room for improvement still, specifically in the areas of the architecture of the network, adjustments to the network's loss function, and refinement of the datasets used for training, making the use of GANs as a denoising tool an enticing avenue for further study.

References

[1] Li Sze Chow and Raveendran Paramesran. Review of medical image quality assessment. *Biomedical Signal Processing and Control*, 27:145–154, May 2016. doi: 10.1016/j.bspc.2016.02.006.

[2] Linwei Fan, Fan Zhang, Hui Fan, and Caiming Zhang. Brief review of image denoising techniques. *Visual Computing for Industry, Biomedicine, and Art*, 2 (1), July 2019. doi: 10.1186/s42492-019-0016-7.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

[4] Amer M. Johri, Vijay Nambi, Tasneem Z. Naqvi, Steven B. Feinstein, Esther S.H. Kim, Margaret M. Park, Harald Becher, and Henrik Sillesen. Recommendations for the assessment of carotid arterial plaque by ultrasound for the characterization of atherosclerosis and evaluation of cardiovascular risk: From the american society of echocardiography. *Journal of the American Society of Echocardiography*, 33(8):917–933, 2020. doi: https://doi.org/10.1016/j.echo.2020.04.021.

[5] P Krishna Kumar, Tadashi Araki, Jeny Rajan, John R Laird, Andrew Nicolaides, and Jasjit S. Suri. State-of-the-art review on automated lumen and adventitial border delineation and its measurements in carotid ultrasound. *Computer Methods and Programs in Biomedicine*, 163:155–168, September 2018. doi: 10.1016/j.cmpb.2018.05.015.

[6] Joseph F. Polak and Daniel H. O'Leary. Carotid intima-media thickness as surrogate for and predictor of CVD. *Global Heart*, 11(3):295, September 2016. doi: 10.1016/j.gheart.2016.08.006.

[7] Jose Seabra and Joao Sanches. Modeling log-compressed ultrasound images for radio frequency signal recovery. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, August 2008. doi: 10.1109/iembs.2008.4649181.

[8] José Carlos Rosa Seabra. *Medical Ultrasound B-Mode Modeling, Despeckling and Tissue Characterization Assessing the Atherosclerotic Disease*. PhD thesis, Instituto Superior Técnico, 2011.

[9] R.F. Wagner, S.W. Smith, J.M. Sandrik, and H. Lopez. Statistics of speckle in ultrasound b-scans. *IEEE Transactions on Sonics and Ultrasonics*, 30(3): 156–163, May 1983. doi: 10.1109/t-su.1983.31404.

Archaea Taxonomic Classification

Jorge Miguel Silva[1]
jorge.miguel.ferreira.silva@ua.pt
Diogo Pratas[1]
pratas@ua.pt
Tânia Caetano[2]
tcaetano@ua.pt
Sérgio Matos[1]
aleixomatos@ua.pt

[1] DETI/IEETA
University of Aveiro
Portugal

[2] CESAM and Department of Biology
University of Aveiro
Portugal

[3] Department of Virology
University of Helsinki
Finland

Abstract

Archaea are a domain of single-celled organisms that live in almost every environment and play significant environmental roles, such as carbon fixation and nitrogen cycling. However, their classification is difficult because most have not been isolated in a laboratory and detected only by their gene sequences in environmental samples. Moreover, archaeal genomes are characterized by significant dissimilarity. This manuscript provides an automatic classification methodology by applying an ensemble method using a combination of reference-free compression measures with GC-content and length. Notably, the results show that we can automatically and accurately distinguish between Archaea genomes at different taxonomic levels.

1 Introduction

Archaea are a domain of single-celled organisms that lack a nucleus. Their cells have unique properties which are distinct from both bacteria and eukaryota domains. Archaea and bacteria are generally similar in size and shape. However, despite the morphological similarities to bacteria, Archaea have genes and metabolic pathways more closely related to eukaryotes, prominently for the enzymes involved in transcription and translation. In addition, other aspects of archaeal biochemistry are unique, such as their reliance on ether lipids in their cell membranes. Furthermore, Archaea are characterized for having a significant genomic inter-dissimilarity.

Despite being firstly detected living in extreme environments such as hot springs and salt lakes with no other organisms, they live in almost every environment. In the human microbiome, they are essential in the gut, mouth, and skin. Furthermore, they play significant environmental roles, such as carbon fixation, nitrogen cycling, organic compound turnover, and maintaining microbial symbiotic and syntrophic communities.

Currently, Archaea are further divided into multiple recognized phyla. However, classification is difficult because most have not been isolated in a laboratory and detected only by their gene sequences in environmental samples. Studying a DNA sequence's complexity (or quantity of information) may help solve this classification problem. As such, this manuscript proposes an Archaea genomic taxonomic classification tool. Specifically, it performs classification without resorting directly to the sequence of the reference genomes. Instead, it uses an ensemble of three predictors, namely normalized compression and two simple property characteristics, for probabilistic classification of DNA sequences.

It is counter-intuitive to think that it is possible to classify a genome recurring only to how much it can be compressed, its length, and the percentage of Guanine and Cytosine. For example, to determine its phylum, order, class or genus. Furthermore, this manuscript shows that it is not only possible but that it can be done automatically with high accuracy, using a small and diverse dataset recurring to alignment-free approaches [11]. The complete study can be fully replicated using the repository https://github.com/jorgeMFS/Archaea.

2 Methods

2.1 Database

The Archaea NCBI database is minimal when compared to other domains of life. The dataset comprises 216 complete reference genomes retrieved from the NCBI database (link) on 30 September 2021. In addition, the taxonomic description was also retrieved from the NCBI database and manually corrected to classify different taxonomic levels correctly. This mapping is available in the project to simplify future usage and replication.

2.2 Normalized Compression (NC)

An efficient compressor, $C(x)$, provides an upper bound approximation for the Kolmogorov complexity ($K(x)$), where $K(x) < C(x) \leq |x|$ ($|x|$ is the length of string x in the appropriate scale). Usually, an efficient data compressor is a program that approximates both probabilistic and algorithmic sources using affordable computational resources (time and memory). Although the algorithmic nature may be more complex to model, data compressors can have embedded sub-programs to handle this nature. The normalized version, known as the Normalized Compression (NC), is defined by

$$\mathrm{NC}(x) = \frac{C(x)}{|x| \log_2 |\mathrm{A}|}, \tag{1}$$

where $C(x)$ is the compressed size of x in bits, $|\mathrm{A}|$ the number of different elements in x (size of the alphabet). Given the normalization, the NC enables to compare the proportions of information contained in the strings independently from their sizes [7]. If the compressor is efficient, then it can approximate the quantity of probabilistic-algorithmic information in data using affordable computational resources. In our work, to determine the NC, we made use of the state-of-the-art DNA sequence compressor: GeCo3 [10].

2.3 Other Measures

The two other measures used to perform Archaea taxonomic classification are the GC-Content (GC) and the length of the genome $|x|$.

GC-Content (GC) represents the proportion of guanine (G) and cytosine (C) bases out the quaternary alphabet $\{A,C,G,T/U\}$. This includes thymine (T) in DNA and uracil (U) in RNA. The GC percentage is given by the number of cytosine (C) and guanine (G) bases in an Archaea genome x with length $|x|$ according to

$$\mathcal{GC}(x) = \frac{100}{|x|} \sum_{i=1}^{|x|} \mathcal{N}(x_i | x_i \in \Xi), \tag{2}$$

where x_i is each symbol of x (assuming causal order), Ξ is a subset of the genomic alphabet containing the symbols $\{G,C\}$ and $\mathcal{N}$ the program that counts the numbers of symbols in Ξ.

GC-content is variable between different organisms. In addition, the GC-content value correlates with the organism's life-history traits, genome size [9], and GC-biased gene conversion [3]. As such, this measure is useful to perform Archaea classification. Furthermore, an organism with a genome high in GC-content is rich in energy and more prone to mutation. Thus, over time, a species tends to decrease its GC-content to become more stable, giving us further information regarding Archaea characterization.

For comparison of the obtained results, we assessed the outcomes obtained using a random classifier. For that purpose, for each task, we determined the probability of a random sequence being correctly classified (p_{hit}) as

$$p_{hit} = \sum_{i=0}^{n} [p(c_i) * p_{correct}(c_i)], \tag{3}$$

where $p(c_i)$ is the probability of each class, determined as

Table 1: Results obtained for Archaea taxonomic classification task regarding the phylum, class, order, family, and genus. The features used were the genome's sequence length (SL), the GC-content (GC) and the Normalized Compression (NC) values. The classifiers used were Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and XGBoost classifier (XGB). The performance was measured using the accuracy (ACC) and the Weighted F1-score (F1-score). The probability of a sequence being correctly classified using a random classifier was determined (p_{hit}).

Data Characteristics			**Random**	**GNB$_{SQ+GC+NC}$**		**SVM$_{SQ+GC+NC}$**		**KNN$_{SQ+GC+NC}$**		**LDA$_{SQ+GC+NC}$**		**XGB$_{NC}$**		**XGB$_{SQ+GC+NC}$**	
Classification	**N. Classes**	**Samples**	p_{hit}	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
Phylum	4	215	25.00	58.14	0.56	68.60	0.56	69.77	0.65	66.28	0.58	77.91	0.76	**80.23**	**0.80**
Class	9	168	11.11	63.24	0.57	61.76	0.52	63.24	0.59	64.71	0.63	66.18	0.65	**69.12**	**0.68**
Order	25	242	4.00	26.80	0.17	28.87	0.16	28.87	0.19	36.08	0.3	34.02	0.34	**41.24**	**0.43**
Family	39	219	2.56	27.27	0.22	29.55	0.15	32.95	0.18	36.36	0.32	35.23	0.34	**39.77**	**0.42**
Genus	97	213	1.03	12.79	0.09	17.44	0.05	15.12	0.04	27.91	0.23	19.77	0.14	**29.07**	**0.24**

$$p(c_i) = \frac{|samples_{class}|}{|samples_{total}|}.$$

On the other hand, $p_{correct}(c_i)$ is the probability of that class being correctly classified. In the case of a random classifier,

$$p_{correct}(c_i) = \frac{1}{|classes|}.$$

3 Results

In this section, we performed five different classification tasks for each Archaea sequence from the dataset. Specifically, the sequences were classified regarding their phylum, class, order, family, and genus.

We applied 5 types of classifiers: Linear Discriminant Analysis (LDA) [6], Gaussian Naive Bayes (GNB) [8], K-Nearest Neighbors (KNN) [4], Support Vector Machine (SVM) [2] and XGBoost classifier (XGB)[1]. To select the best performing method we computed the Accuracy and the Weighted F1-score.

Furthermore, we performed classification using three different features: the Normalized Compression (NC), GC-content (GC), and sequence length (SL). These three features were fed to all the classifiers, and the accuracy and weighted F1-score were measured to determine which classifier was best suited for this task.

Table 1 depicts the accuracy and weighted F1-score values obtained for each classifier. For all classification tasks, the best performing classifier was the XGBoost classifier. Regarding the features used, despite the NC feature being the most relevant, combining it with the GC-Content and Sequence Length improved the accuracy and F1-score result. This improvement increased when the number of classes was higher. Overall, there is a decrease in accuracy and F1-score when there is an increase in the number of classes. Specifically, we obtained the best performance in the phylum classification of the Archaea (accuracy - 80.23%, F1-score - 0.80) and our lowest performance in genus classification (accuracy - 29.07%, F1-score - 0.24). This decrease is mainly because the average number of samples per class decreases as the number of classes increases. As such, many classes lack a valid number of samples to be accurately classified. Moreover, part of the classification inaccuracies can be explained by possible errors in the assembly process of the original sequence or eventual sub-sequence contamination of parts of the genomes. Other inaccuracies could be due to several genomes being reconstructed using older methods that have been improved since then [5].

As far as we know, this is the first attempt at performing this type of taxonomic classification using reference-free methods. As such, for comparison purposes, we assessed the outcomes obtained using a random classifier. Specifically, for each task, we determined the probability of a random sequence being correctly classified (p_{hit}). Overall, there is a vast improvement relative to the random classifier, showing the importance of the features used in the classification process. The results are particularly encouraging given the small sample size and the many classification labels of the dataset. We conclude that these classification results show that this metric can be utilized for taxonomic classification, particularly if more sequence samples are added to the public dataset.

4 Conclusion

This manuscript evaluates the capability of using complexity measures to perform Archaea classification at different taxonomic levels. For this purpose, we used the NC, the GC-content, and length of the genome sequence. The best results were obtained using all mentioned features in the XGBoost classifier. Notably, the results showed that we can automatically and accurately distinguish between Archaea genomes at different taxonomic levels. As far as we are aware, this is the first study where reference-free classification of the Archaea at different taxonomic levels is performed. As such, we compared our obtained results with a random classifier. As a result, we extensively outperform a random classifier, proving these measures' efficiency in performing this type of classification. However, the results obtained showed a decrease in accuracy when approaching the lowest taxonomic levels due to an increase in the number of classes and a decrease in the number of samples per class. As such, when future entries are added to the database, accuracy may significantly increase in the lowest taxonomic levels. Future work involves the addition of other experts regarding the proteomes of Archaea.

Overall, this manuscript shows that the efficient approximation of the Kolmogorov complexities of Archaea sequences as measures of complexity have a profound impact on genomes identification and classification.

References

[1] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2.

[2] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[3] Laurent Duret and Nicolas Galtier. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, 10:285–311, 2009.

[4] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. *Springer Berlin Heidelberg*, pages 986–996, 2003. doi: 10.1007/978-3-540-39964-3_62.

[5] Jennifer Lu and Steven L Salzberg. Removing contaminants from databases of draft genomes. *PLoS computational biology*, 14(6): e1006277, 2018.

[6] Geoffrey J McLachlan. *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons, 2004.

[7] Diogo Pratas and Armando J Pinho. On the approximation of the Kolmogorov complexity for DNA sequences. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 259–266. Springer, 2017.

[8] Irina Rish et al. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[9] Jonathan Romiguier, Vincent Ranwez, Emmanuel JP Douzery, and Nicolas Galtier. Contrasting gc-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome research*, 20(8):1001–1009, 2010.

[10] Milton Silva, Diogo Pratas, and Armando J Pinho. Efficient DNA sequence compression with neural networks. *GigaScience*, 9(11), 11 2020. ISSN 2047-217X. giaa119.

[11] et al. Zielezinski, Andrzej. Benchmarking of alignment-free sequence comparison methods. *Genome Biology*, 20(144), 2019. doi: 10.1186/s13059-019-1755-7.

Author Index