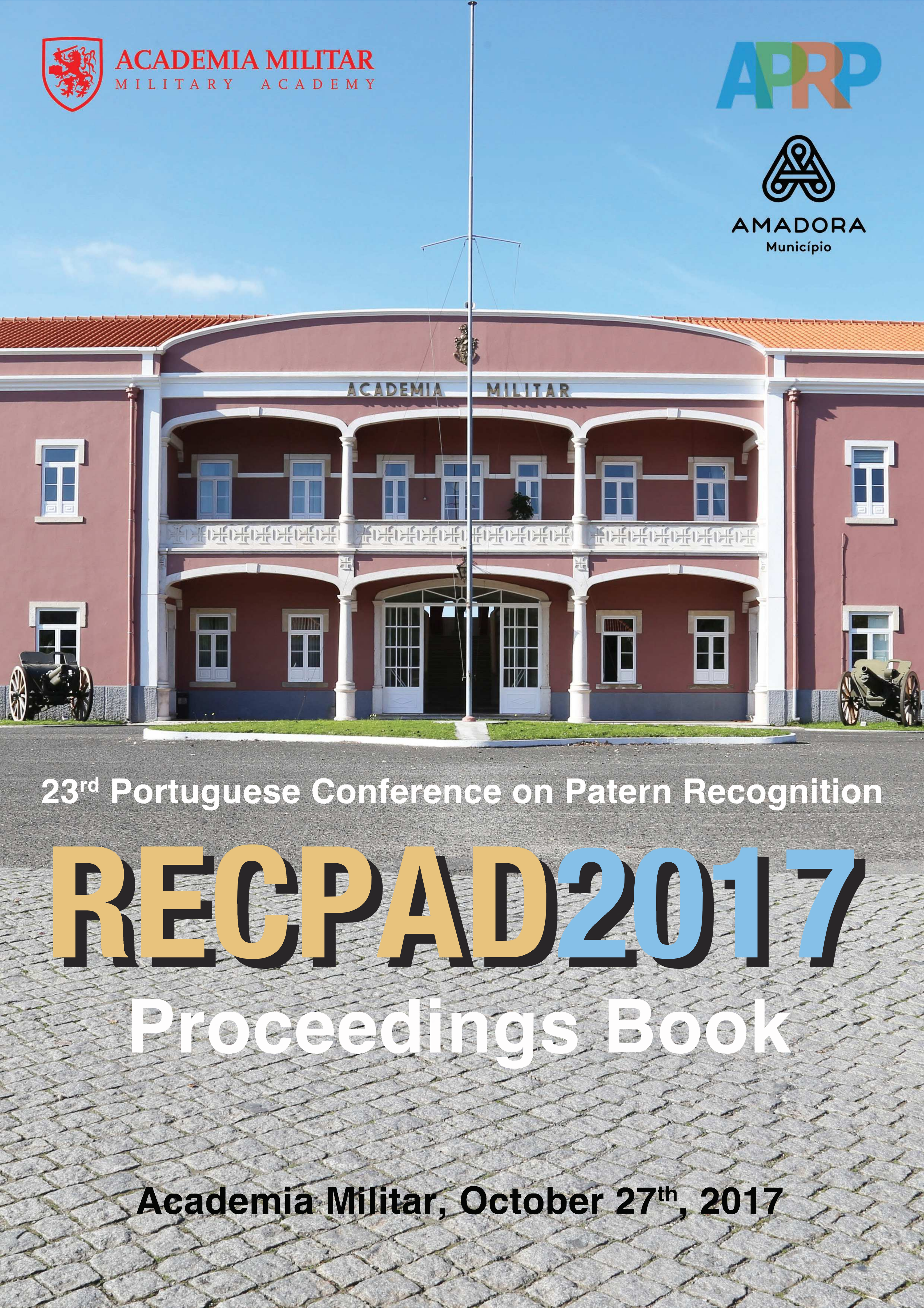
ACADEMIA MILITAR
MILITARY ACADEMY

APRP

AMADORA
Município



23rd Portuguese Conference on Patern Recognition

# RECPAD2017

## Proceedings Book

Academia Militar, October 27th, 2017

# RECPAD 2017

23rd Portuguese Conference on Pattern Recognition

**Amadora, Portugal**

**October 27th, 2017**

## Organizing Committee

José Serra da Silva, Chairman

Jorge Torres

Pedro Mendonça dos Santos

Thomas Gasche

## Local support

Commandant's Cabinet (AM)

SIAM – Secção de Informática da Academia Militar

DSA - Departamento Serviços Académicos (AM)

## Organised by:

Portuguese Military Academy

## Sponsors:

APRP – Associação Portuguesa de Reconhecimento de Padrões

CMA – Câmara Municipal da Amadora

# Table of Contents

## Scientific Committee

Alexandre Bernardino (IST)
Ana Fred (IST)
Ana Aguiar (FEUP)
Ana Maria Mendonça (FEUP)
Ana Maria Tomé (UA)
André Marçal (FCUP)
Andrzej Wichert (IST)
António Neves (UA)
António Pinheiro (UBI)
Armando Pinho (UA)
Augusto Silva (UA)
Aurélio Campilho (FEUP)
Beatriz Sousa Santos (UA)
Bernardete Ribeiro (UC)
Catarina Silva (IPL)
Fernando Monteiro (IPB)
Hans du Buf (UAlg)
Helder Araújo (UC)
Hélder Oliveira (FCUP)
Hugo Proença (UBI)
Jaime Cardoso (FEUP)
Jaime Santos (UC)
Joao Barreto (UC)
João Barroso (UTAD)
João Cardoso (UC)
João Rodrigues (UAlg)
João Sanches (IST)
João Tavares (FEUP)
Joaquim Pinto da Costa (FCUP)
Jorge Barbosa (FEUP)
Jorge Batista (UC)
Jorge S. Marques (IST)
Jorge Santos (ISEP)
Jorge Torres (Acad Militar)
José Bioucas-Dias (IST)
José Silva (Acad Militar)
Luís A. Alexandre (UBI)
Luís F. Teixeira (FEUP)
Mário Figueiredo (IST)
Miguel Coimbra (FCUP)
Miguel Correia (IST)
Noel Lopes (IPG)
Nuno Martins (ISEC)
Paulo Carvalho (UC)
Paulo Oliveira (IST)
Paulo Salgado (UTAD)
Pedro Pina (IST)
Ricardo Morla (FEUP)
Rodrigo Ventura (IST)
Samuel Silva (UA)
Susana Vinga (IST)
Thomas Gasche (Acad Militar)
Verónica Vasconcelos (ISEC)

*Academia Militar*

# *Editorial*

*The Military Academy, a Military Institution of Higher Education, is pleased to host the 23rd Portuguese Conference on Pattern Recognition on the 27th of October, 2017.*

*I welcome the invited speaker Prof. Paul Scheunders from Antwerp University and also participants have travelled from the four corners of Portugal; from the Algarve, Trás-os-Montes, the western coastline and the most interior of all our universities. I welcome all of you to the Military Academy confident that the day that you spend here will be constructive*

*The Military Academy, which has produced Officers for the army since 1790 in its original designation of "Academia Real de Fortificação, Artilharia e Desenho" has always had a strong academic component in its preparation of its cadets. This continues today with the support that the Military Academy Centre for Research and Development, CINAMIL, provides both for our own staff and also external projects, permitting growth in the collaboration between academics, the military and industry in many areas of mutual interest.*

*The number of presentations and participants shows the importance that the national community gives to pattern recognition and the number of different areas where this is applied, including such questions as the classification of aerial images and the detection of mines/explosives.*

*I would like to thank the Portuguese Association of Pattern Recognition for having chosen our Academy to host this year's meeting and also to thank all those involved in the organization of this meeting, including the Organization Committee.*

*Finally, I would like to thank all that are present at this meeting for their participation today and I am sure that we all gain from this possibility to share ideas, methods, techniques and results.*

*The Commander of the Military Academy*

*João Jorge Botelho Vieira Borges*
*Major General*

Published works are the sole responsibility of their authors.

# Invited Speaker

# Machine Learning for Remote Sensing Image Analysis

Paul Scheunders
http://www.visielab.uantwerpen.be/scheunders

Vision Lab, Department of Physics
University of Antwerp, campus CDE
Belgium

## Abstract

In this talk, I will describe the state of the art on the development and application of machine learning methodologies in the remote sensing domain. I will describe the specific remote sensing analysis problems that are typically handled by machine learning. On important type of sensor is the hyperspectral image sensor. Hyperspectral images contain many spectral bands, each revealing the earth surface reflected light at a particular wavelength. Hyperspectral image sensors have been important tools for the characterization of materials based on their light reflectance mainly in remote sensing but in other domains as well. These hyperspectral images require specific image processing and analysis methodologies.

- By far the most prominent task is *land cover mapping*. Here, each pixel is to be assigned to one of a number of predefined land cover classes (e.g. water, sand, vegetation, ...). Mostly, this is done using the supervised approach, where labeled training samples are obtained e.g. by ground campaigns. Land cover maps can be obtained from any type of remote sensing data [5].

- Another important task is the *estimation of land physical parameters* from the data. Here, a forward model describes a measurement in function of a number of parameters (e.g. moisture) and measurement conditions (e.g. viewing direction). Model inversion then tries to estimate the physical parameters from the measurements. Besides a long tradition of inversion methods based on look up tables using spectra generated by physical models, supervised methods are nowadays being used to estimate these parameters using a training set of input (spectra)-output (parameter values) data samples. In this case, the output labels are continuous variables, which requires regression methods [15].

- Satellite sensor platforms fly over the same regions with varying revisiting periods, which allows to study changes of the earth surface. In particular, multitemporal *change detection* is an important research domain [4]. Applications can vary from seasonal vegetation mapping to disaster monitoring of floods, tsunami's etc. The problem of change detection can be treated in a supervised or unsupervised manner. One particular problem that has to be tackled is that light reflectance of materials heavily depends on acquisition conditions and external factors, which may vary over time. To tackle such specific problems, researchers have been considering strategies based either on physical models or on the machine learning field of *domain adaptation* [14].

- A particular property of many remote sensing data is the high dimensionality. A pixel may contain many measurements as in hyperspectral images, or may be represented by a large number of features, e.g. contextual features from the pixel's neighborhood. Special attention is given to the particular problem of *feature selection and extraction*, or *dimensionality reduction* [2]. Natural images are in general spatially smooth, remote sensing images make no exception. The research area of *spectral-spatial classification* makes advantage of this property to enhance the performance of classification algorithms [8].

- The ever increasing number of remote sensing products allows to combine information from different sensors to obtain more information than can be obtained from individual sensors. One example is the combination of information from a high spatial resolution sensor with a high spectral resolution sensor. Such combinations can improve classification performances [7, 9]. The information can be fused at the image level (*image fusion*) or at the level of the classifier (*decision fusion*). Another emerging task due to the ever increasing amount of remote sensing data is the *retrieval of images* from large databases [6, 13].

- One particular analysis problem is the detection of targets smaller than the pixel size. In *target detection*, one searches for particular materials with a size smaller than the pixel size [12]. If the spectral signature of the target is not known, one refers to anomaly detection [11]. A very active field of research is *spectral unmixing*, where one wants to discover the fractional abundance of materials within a pixel [3]. These materials can be mixed because of the limited spatial resolution of a pixel, or can be intimately mixed at microscales. Finally *subpixel mapping* aims at obtaining land cover classification maps at the subpixel level, e.g. by combining unmixing with a spatial arrangement of the obtained abundances within a pixel [1]. *Superresolution* methods on the other hand try to improve the spatial resolution of the images before analysis [10].

In the relevant literature of the last few years, specific analysis methods have been developed for these remote sensing tasks. Most of the developed methods apply machine learning methodologies, which can be divided into 6 distinct groups:

1. *kernel methods*, where models are based on data similarity and projections into higher dimensional spaces;

2. *neural network methods*, where the input-outputs relations are learned by backpropagating errors through a series of operations learned from data;

3. *manifold learning methods*, where the underlying nonlinear structure of the data is taken into account by the model;

4. *structured output learning methods*, where prior knowledge about the problem to be solved is encoded as structural relationships among the outputs (e.g. in the spatial domain);

5. *ensemble learning methods*, where, instead of using a single method for prediction, many imperfect models are run and then their average is used as a solution able to generalize better;

6. *sparse learning methods*, where models reduce the dependency on specific training data or features by selecting subsets that are relevant to the problem.

In the talk, a section is devoted to each of these groups. Each section contains a short introduction with basic notions and methods, a review of the recent state of the art, with a conceptual description of some more involved methods.

## References

[1] Peter M. Atkinson, Eulogio Pardo-Iguzquiza, and Mario Chica-Olmo. Downscaling cokriging for super-resolution mapping of continua in remotely sensed images. *IEEE Trans. Geosci. Remote Sens.*, 46(2):573–580, Febr. 2008.

[2] JA Benediktsson, M Pesaresi, and K Arnason. Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. *IEEE Trans. Geosci. Remote Sens.*, 41(9, 1):1940–1949, Sep. 2003. ISSN 0196-2892. doi: {10.1109/TGRS.2003.814625}.

[3] Jose M. Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based Approaches. *IEEE J. Sel. Topics Appl. Earth Observ. and Remote Sens.*, 5(2, SI):354–379, 2012.

[4] Lorenzo Bruzzone and Francesca Bovolo. A Novel Framework for the Design of Change-Detection Systems for Very-High-Resolution Remote Sensing Images. *P. IEEE*, 101(3, SI):609–630, March 2013.

[5] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson. Advances in hyperspectral image classification. *IEEE Signal Proc. Mag.*, 31:45–54, 2014.

[6] M Datcu, K Seidel, and M Walessa. Spatial information retrieval from remote-sensing images - Part 1: Information theoretical perspective. *IEEE Trans. Geosci. Remote Sens.*, 36(5, 1):1431–1445, Sep. 1998.

[7] Mathieu Fauvel, Jocelyn Chanussot, and Jon Atli Benediktsson. Decision fusion for the classification of urban remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 44(10, 1):2828–2838, Oct. 2006.

[8] Mathieu Fauvel, Yuliya Tarabalka, Jon Atli Benediktsson, Jocelyn Chanussot, and James C. Tilton. Advances in Spectral-Spatial Classification of Hyperspectral Images. *P. IEEE*, 101(3, SI):652–675, 2013.

[9] L. Gomez-Chova, D. Tuia, G. Moser, and G. Camps-Valls. Multimodal classification of remote sensing images: A review and future directions. *P. IEEE*, 103(9):1560–1584, Sept 2015. ISSN 0018-9219.

[10] Yanfeng Gu, Ye Zhang, and Junping Zhang. Integration of spatial-spectral information for resolution enhancement in hyperspectral images. *IEEE Trans. Geosci. Remote Sens.*, 46(5):1347–1358, May 2008.

[11] Dimitris Manolakis, Eric Truslow, Michael Pieper, Thomas Cooley, and Michael Brueggeman. Detection Algorithms in Hyperspectral Imaging Systems. *IEEE Sign. Process. Mag.*, 31(1):24–33, Jan. 2014.

[12] Nasser M. Nasrabadi. Hyperspectral Target Detection. *IEEE Sign. Process. Mag.*, 31(1):34–44, 2014.

[13] M Schroder, H Rehrauer, K Seidel, and M Datcu. Spatial information retrieval from remote-sensing images - Part II: Gibbs-Markov random fields. *IEEE Trans. Geosci. Remote Sens.*, 36(5, 1):1446–1455, Sep. 1998.

[14] D. Tuia, C. Persello, and L. Bruzzone. Recent advances in domain adaptation for the classification of remote sensing data. *IEEE Geosci. Remote Sens. Mag.*, 4(2):41–57, 2016.

[15] Jochem Verrelst, Luis Alonso, Gustavo Camps-Valls, Jesus Delegido, and Jose Moreno. Retrieval of Vegetation Biophysical Parameters Using Gaussian Process Techniques. *IEEE Trans. Geosci. Remote Sens.*, 50(5, 2):1832–1843, May 2012.

# Poster Session I

# Enabling MIMO Beamforming through compressed CSI feedback based on Principal Component Analysis

Jessica Sanson
jessikbs.37@gmail.com
Pedro Tome
tome.p.m@ua.pt
Petia Georgieva
petia@ua.pt

University of Aveiro, DETI
Aveiro, Portugal

## Abstract

Multiple-input multiple-output (MIMO) wireless systems provide a mechanism to increase the reliability of signal reception: beamforming. Beamforming requires knowledge of the wireless channel by the transmitter, and this channel state information (CSI) is usually provided by the receiver through a feedback mechanism. When the total number of antennas is large, the CSI matrix becomes too large to be fed back. In this paper we propose two methods for the compression of the CSI matrix based on principal component analysis (PCA).

## 1 Introduction

MIMO systems are increasingly adopted in communication systems due to the potential capacity gains they offer when using multiple antennas. Multiple antennas use the spatial dimension, beyond time and frequency, without changing the system bandwidth requirements. MIMO approach was adopted in recent wireless standards, such as the 802.11x families, to provide much higher data rates. Because MIMO uses antenna arrays, beamforming — the focusing of the power radiated by a radio transmitter into thin, steerable beams — can be adopted to improve the received signal-to-noise ratio (SNR) which, in turn, reduces the bit error rate (BER).

In order for the optimal instantaneous beam to be formed, the transmitter requires full knowledge of the wireless channel. This is usually done by feeding back a channel state information (CSI) matrix from the receiver to the transmitter. In practice, the overhead incurred to obtain the (potentially enormous) CSI matrix can use most of the available system resources, rendering this feedback mechanism impractical. To solve this problem, the authors of [3, 4] use statistical information feedback — for example, the channel mean or channel covariance matrices. Other approaches, described in [1, 2], consist in limiting the rate at which CSI feedback is performed. However, these methods generally do not work as well as those that use instantaneous feedback on each transmission, since they do not follow the rapid fluctuations of the channel. In this paper we propose a solution to this problem based on compression-decompression of the CSI matrix following the PCA approach.

## 2 Proposed Method

We consider a MIMO system with a base station with $N_T$ transmitting antennas and a mobile receiver with $N_R$ receiving antennas, where $N_T \geq N_R$. Communication is frequency division multiplexed between $N_f$ subcarriers. The CSI matrix is denoted as $\mathbf{H}(f) \in \mathbf{C}^{N_R \times N_T}$. The Beamforming aims to focus the power radiated by a radio transmitter into thin, steerable beams in order to improve the SNR of a given transmission. The CSI matrix is used by the transmitter in order to steer the beam in the direction which maximizes the power delivered to the receiver. To obtain this channel information, the beamformer (the transmitter) sends a channel sounding packet to the beamformee (the receiver), which estimates the CSI matrix based on how the channel distorted the sounding packet. Finally, the beamformee feeds back the CSI to the beamformer, in order to create a steering matrix for the data transmission beam.

The proposed algorithm consists of two distinct processes: compression of the CSI matrix at the receiver, and subsequent restoration (decompression) at the base station.

### 2.1 CSI compression at the receiver

After the receiver receives the channel sounding packet from the base station, it performs a series of measurements in order to estimate the channel and calculate the CSI matrix $\mathbf{H}(f)$. Then, PCA is applied by calculating the covariance matrix of $\mathbf{H}$ and then applying eigendecomposition to it, $cov(\mathbf{H}) = \mathbf{V}\Lambda\mathbf{V}^H$. The compression matrix $\mathbf{U}_c$ is the eigenvector matrix $\mathbf{V}$ with the removal of the vectors related to the less significant eigenvalues. The selection of the $k$ most significant eigenvalues is performed according to a chosen Eigen Ratio (ER), which is a configuration parameter of the algorithm. The compressed CSI matrix is $\mathbf{H}_c = \mathbf{H}\mathbf{U}_c$

### 2.2 CSI decompression at the base station

After receiving the compressed CSI information from the receiver, the base station must decompress it and restore the CSI matrix. Two approaches are studied:

1) **CSI feedback with $\mathbf{U}_c$.**

The receiver feeds back to the base station both $\mathbf{H}_c$ and $\mathbf{U}_c$ matrices. The decompressed CSI matrix is then $\mathbf{H}_d = \mathbf{H}_c\mathbf{U}_c^H$

2) **CSI feedback without $\mathbf{U}_c$.**

The receiver feeds back to the base station only $\mathbf{H}_c$ matrix. The base station computes $\mathbf{U}_c$ based on past CSI data. Assuming that, at some past time instant $t - \Delta t$, the base station had a CSI matrix $\mathbf{H}_d(t - \Delta t)$, then for the present time instant $t$, $\mathbf{U}_c$ is computed from the eigen-decomposition $cov(\mathbf{H}_d(t - \Delta t)) = \mathbf{V}\Lambda\mathbf{V}^H$. $k$ is equal to the number of columns of the received $\mathbf{H}_c$ matrix. We assume that $\mathbf{H}_d(t - \Delta t)$ matrix corresponds to the last $\mathbf{H}_d$ matrix calculated by the base station. Since there is no previous $\mathbf{H}_d$ matrix for the transmission of the first packet, the CSI feedback without $\mathbf{U}_c$ runs after at least one data packet has been transmitted.

## 3 Simulation Results

In order to evaluate the proposed CSI feedback matrix compression algorithms, we simulated a MIMO IEEE 802.11ac wireless LAN communication system where VHT packets were transmitted over two different specifications of the TGac channel models: "Model-B", which (among other things), has a maximum delay of 80 ns, and "Model-C", which has a maximum delay of 200 ns. The other simulation parameters are given in Table 1.

The performance measures are the achieved BER as a function of the SNR and the compression ratio ($CR$), defined in (1). For the CSI feedback with $\mathbf{U}_c$, the size of the compressed CSI is the sum of the number of elements of $\mathbf{H}_c$ and $\mathbf{U}_c$ matrices. For the CSI feedback without $\mathbf{U}_c$, the size of the compressed CSI is the number of elements of $\mathbf{H}_c$ matrix. Lower the value of $CR$, higher is the compression of the CSI matrix.

$$CR = \frac{\text{Size of compressed CSI}}{\text{Size of uncompressed CSI}} \tag{1}$$

Table 1: Simulation Parameters

| PARAMETER | VALUE |
|---|---|
| Number of Transmitting Antennas ($N_T$) | 8 |
| Number of Receiving Antennas ($N_R$) | 4 |
| Number of Space-Time Streams | 2 |
| TGac Channel Model | B (80 ns) and C (200 ns) |
| Channel Bandwidth | 80 MHz |
| Number of Occupied Subcarriers | 242 out of 256 |
| Modulation Coding Scheme | 256-QAM rate-5/6 |
| Eigen Ratio ($ER$) | 0.65, 0.80, 0.90 and 0.95 |

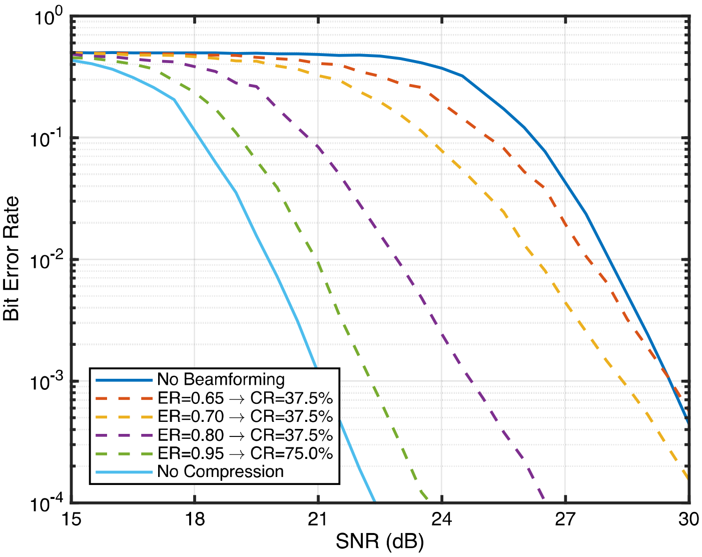


Figure 1: BER vs. SNR vs. *ER* results for the transmission with CSI feedback with $\mathbf{U}_c$, in a Model-B TGac channel.

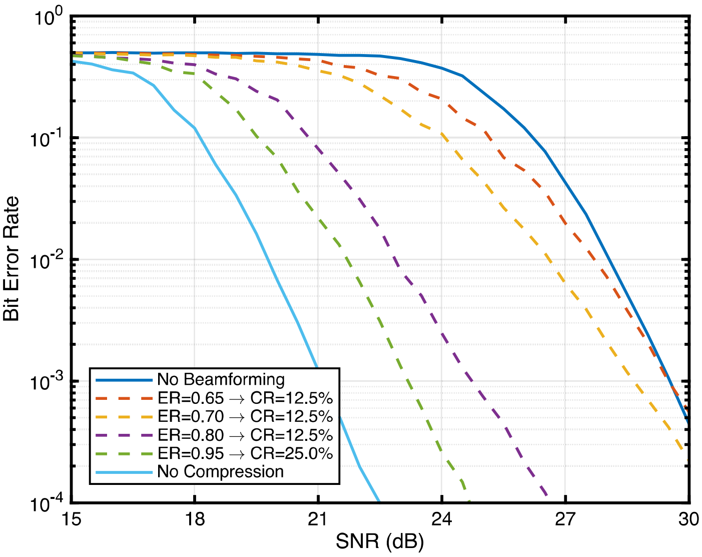


Figure 2: BER vs. SNR vs. *ER* results for the transmission with CSI feedback without $\mathbf{U}_c$ in a Model-B TGac channel.

For reference, we compared our results with the cases where there is no CSI feedback (no beamforming) and where the CSI feedback is perfect (no compression).

Fig. 1 illustrates the variation of the BER and the *CR* as a function of the SNR for transmission of CSI feedback with $\mathbf{U}_c$, in a Model-B TGac channel for various values of the Eigen Ratio (*ER*). Lower the BER and the *CR*, better are the results. Fig. 1 also shows the BER curves for the case with no beamforming and the case with no compression of the CSI matrix (i.e. perfect feedback). The figure clearly shows that the proposed algorithm achieves great improvements over the case with no beamforming even with high levels of compression: for instance, for $ER = 0.95$, we only get a BER penalization of just over 1 dB for a compression ratio of 75%, meaning that the compressed CSI is 25% smaller than the uncompressed CSI. Also noteworthy, using $ER = 0.80$ we can decrease the feedback CSI to 37.5% of its uncompressed size in exchange for a 4 dB penalization in BER.

Fig. 2 is similar to the previous figure, except that it relates to the algorithm for the CSI feedback without $\mathbf{U}_c$. It is important to notice that the BER curves are almost identical to those of Fig. 1, yet the resulting compression ratios are much lower. Phenomenally, we can achieve the same BER performance as the previous example of CSI feedback with $\mathbf{U}_c$, for $ER = 0.80$ and still get a 3 times better compression ratio.

In order to get a broader understanding of the pros and cons of the proposed algorithms, we repeated the simulations of the transmission with CSI feedback with and without $\mathbf{U}_c$, for a different channel: the Model-C TGac channel, which has a delay larger than that of the Model-B TGac channel previously used. For the sake of brevity, we chose $ER = 0.80$. Fig. 3 illustrates the results of those simulations. As explained, the BER performance in a Model-B channel is largely independent of the compression algorithm that is used. This is clearly not the general case, however, as the results of the Model-C simulations show: using the algorithm where the $\mathbf{U}_c$ matrix is not fed back to the base station causes a higher BER in the Model-C channel. This is to be expected, since the Model-C channel exhibits faster fading than the Model-B channel (the channel varies more rapidly), which means that the $\mathbf{U}_c$ matrix that is calculated using the CSI from a previous packet is more outdated and is less representative of the current channel state. In spite of that, the BER in either channel is still much lower than that of the transmission without any CSI feedback (no beamforming).

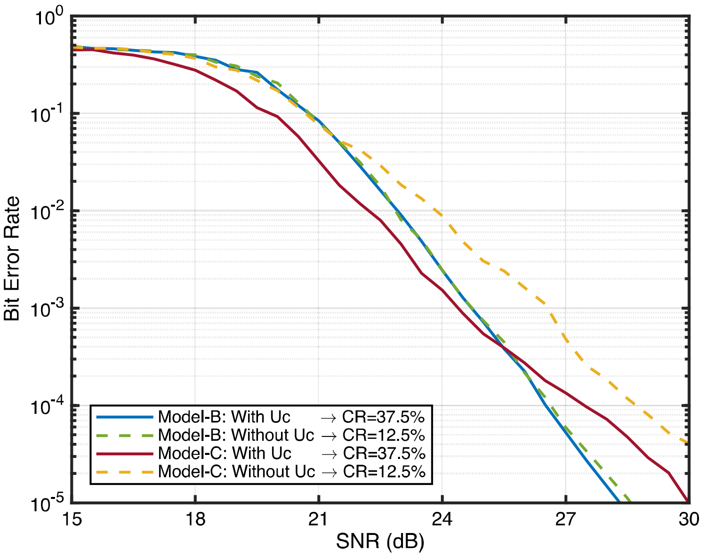


Figure 3: BER vs. SNR results for the transmission with CSI feedback with (full lines) and without (dashed lines) $\mathbf{U}_c$ in a Model-B and Model-C TGac channel. In all cases, $ER = 0.80$.

## 4 Conclusion

In this work we studied two PCA-based methods to compress the feedback CSI of a MIMO communications system. The first method compressed the full CSI matrix $\mathbf{H}$, into two smaller matrices, $\mathbf{H}_c$ and $\mathbf{U}_c$, and the second method compressed the CSI matrix only into $\mathbf{H}_c$ matrix.

Realistic 802.11ac WLAN simulations confirmed that either method provides great benefit over not doing any CSI feedback at all (due to impractically large $\mathbf{H}$ matrices, for instance), and that the achieved compression ratios of the second method are far better than those of the first method — enough, at least, to compensate any reduction in BER performance. Finally, the same simulations revealed that both methods are robust to channel delay.

## References

[1] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang. On beamforming with finite rate feedback in multiple-antenna systems. *IEEE Transactions on Information Theory*, 49(10):2562–2579, October 2003.

[2] M. Trivellato, F. Boccardi, and F. Tosato. User selection schemes for MIMO broadcast channels with limited feedback. In *IEEE Veh. Technol. Conf.*, pages 2089–2093, Dublin, Ireland, April 2007.

[3] E. Visotsky and U. Madhow. Space-time transmit precoding with imperfect feedback. *IEEE Transactions on Information Theory*, 47 (6):2632–2639, September 2001.

[4] S. Zhou and G. B. Giannakis. Optimal transmitter eigen-beamforming and space-time block coding based on channel correlations. *IEEE Transactions on Information Theory*, 49(7):1673–1690, July 2003.

# Importance of the critical sampling size on data analytics

José Silva[1]
jmpsilva@student.dei.ic.pt
Bernardete Ribeiro[1]
bribeiro@dei.uc.pt
Andre H. Sung[2]
andrew.sung@usm.edu
César Teixeira[1]
cteixei@dei.uc.pt

[1] Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal
[2] School of Computing
University of Southern Mississippi
Mississippi, USA

## Abstract

Nowadays, in the Big Data Era, information is generated and shared at an enormous pace. The amount of data flowing, thanks to the Internet of Things has increased immeasurably (Agarwal and Dhar) [1]. Big Data has also become a powerful tool used by different organizations for applications ranging from decision support, predictive and prescriptive analytics, to knowledge extraction and intelligence discovery.
However, because of the large amount of information that is generated and consumed, data first needs to be "filtered" or selected. The tasks of knowledge discovery could be very expensive, such as resources and time consumption, and therefore, it is highly desirable to have techniques that can help selecting only the necessary and sufficient data instead of using entire datasets. In this paper, we present a heuristic method to find the critical sampling of datasets and thus reducing the cost of the data mining tasks. The heuristic yielded satisfactory and promising results.

## 1 Critical Sampling Size

The concept of CSS has been defined by Sung *et al.* [4], as the absolute minimal number of examples, of a given dataset, that are needed for a specific learning machine to achieve some performance threshold. To formally define this problem, a given dataset is represented by $D_n$, where $n$ denotes the number of examples at hand. The learning machine is designated by $M$ and the performance threshold by $T$. With respect to the notation, $v$ is used to denote the CSS where $v \leq n$.
This way, using these notations it is possible to define it with the following two conditions. In order for $v$ to actually represent the Critical Sampling Size of a specific dataset, the following two conditions must hold.

1-There exists a $D_v$ which lets $M$ achieve a performance of at least $T$, where $D_v$ is a sample of $D_n$ with $v$ examples and $P_M(D_v)$ is the performance of $M$ when trained with $D_v$

$$(\exists D_v \subset D_n)[P_M(D_v) \geq T] \tag{1}$$

2-For all $j < v$, a sample of $D_n$ with $j$ examples fails to let $M$ achieve a performance of at least $T$

$$(\forall D_j \subset D_n)[j < v \Rightarrow P_M(D_j) < T] \tag{2}$$

Thus, $v$ is the critical (absolute minimal) number of data examples required in any sampling to ensure that the performance of $M$ meets the given performance threshold $T$.
The reason for the above wording is that (1) the critical sampling depends on three things: $D_n$, $M$, and $T$, (2) multiple critical samplings may exist, (3) critical sampling may not exist at all as well, e.g. when $T$ is too high or the whole dataset is too small to allow the $M$ achieve performance $T$.

The problem of deciding if a given number $v$ is the CSS for the dataset $D_n$, with respect to a given learning machine $M$ and performance threshold $T$ is both NP-hard and coNP-hard; a sketch of the proof is given in Sung *et al.* [4]. For this reason, heuristic methods are used, which can also obtain satisfactory results, as can be seen in related work Suryakumar *et al.* [2], by the same authors, to find the Critical Feature Dimension.

### 1.1 Heuristic Method

The heuristic is a sequential example selection method composed of five steps. It is in fact a meta-heuristic algorithm since it is a higher-level procedure that resorts to other algorithms (learning machines) in order to solve the CSS problem.

This heuristic fits into the Wrapper methods group, since it treats the learning machines as black boxes and uses their performance as the objective function to evaluate the critical sample being created.

The heuristic method is defined by the following steps:

1. Apply a clustering algorithm like k-means to partition $D_n$ into $k$ clusters. ($k$ can be determined, for example, by the number of classes in the dataset)
2. Select, say randomly, $m$ examples from each cluster to form a sample with $m * k$ examples. (The value $m$ is set to be fairly small)
3. Supplement the sample with additional $d * k$ (for some d) examples, selected randomly from the whole dataset $D_n$, to form a sample $D_v$.
4. Apply learning machine $M$ on the sample, then measure the performance $P_M(D_v)$.
5. If $P_M(D_v) \geq T$, then $D_v$ is a critical sampling, and its size $v$ is the Critical Sampling Size for ($D_n$, $M$). Otherwise enlarge $D_v$ by repeating step 2 and step 3 until a critical sampling is found, or until the whole $D_n$ is exhausted and the procedure fails to find $v$.

## 2 Experimental Setup

The datasets used in this work were downloaded from the UCI Machine Learning Repository [3]. Their dimensions vary both in the number of examples and in the number of features. This way, it is possible to better visualize the quality of the heuristic dealing with datasets of different characteristics. Table 1 presents some informations about the used datasets to better understand the differences between them.

Table 1: Datasets characteristics

| | Ads | Credit | Hapt | Isolet |
|---|---|---|---|---|
| **# Features** | 1558 | 23 | 561 | 617 |
| **# Classes** | 2 | 2 | 12 | 26 |
| **# Examples** | 3279 | 30000 | 10929 | 7797 |

### 2.1 Sampling Method

The size of both the training and test sets can influence the performance, that a learning machine achieves. This is an important aspect thus, when comparing the performance of a data mining task using the whole dataset and the sampled data, this must be taken in consideration. Because of this, three ways to split the datasets were considered, and therefore, for each way, a respective CSS will be heuristically determined. Each method consists in splitting the dataset in different ratios, which are:

- 30% Train / 70% Test
- 50% Train / 50% Test
- 70% Train / 30% Test

For instance, when using the 30% for training and 70% for testing, to compare the results, the critical sample, constructed by the heuristic, is also tested with 70% of the dataset.

As stated in Section 1.1, the CSS is obtained iteratively. First the data is clustered using k-means and then, from each cluster, $m$ examples are selected to form the sample $Dv$. To analyze the usefulness of this method, another two approaches were used.

- $mk+r$: Initial approach, where the sample $D_v$ is composed by selecting $m$ examples from each one of the $k$ clusters and then complemented with more $d*r$ random examples.
- $mk$: Same as above except that the sample $D_v$ is not complemented with $d*$ random examples.
- $r$: Random sampling. The construction of $D_v$ is made by randomly selected examples. In order to maintain some consistency, here $r$ can be calculated with $m*k$.

Regarding the way as the examples are sampled from each cluster, 3 ways were studied, (Asc) by ascending order of distance to the cluster centroid, (Decr) by decreasing order, and (Rand) randomly.

### 2.2 Parameters Setup

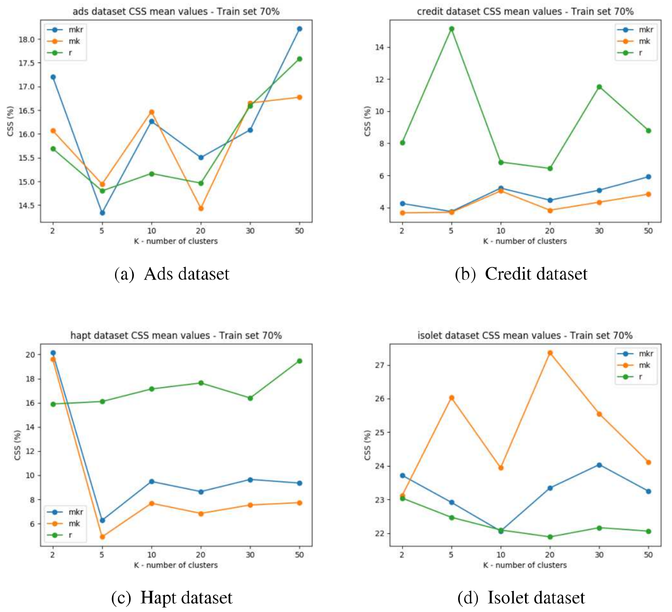
The proposed heuristic to find the CSS has four parameters (see 1.1). Their values should be decided by considering $(i)$ the nature of the problem, $(ii)$ the size of the dataset, $(iii)$ the data mining task that will be applied and $(iv)$ the amount of available resources. These four parameters are, respectively, $k$, $m$, $d$ and $T$ and Table 2 shows the values that were used for them.

Table 2: Settings for the heuristic parameters

| Parameter | $k$ | $m$ | $d$ | $T$ |
|---|---|---|---|---|
| **Description** | # of clusters | # of instances from each cluster | # of instances $(*k)$ to supplement sample | threshold value |
| **Used value** | 2 , 5, 10, 20, 30, 50 | 1% of cluster | $\frac{m}{4}$ | $P_M(D_n)$ |

## 3 Results and Discussion

Each dot of the lines represents the averaged CSS value, of 30 runs, for each value of $k$. For the $mk+r$ and $mk$ sampling methods, the value that is represented is the one that obtained the best results among the *Asc*, *Decr* and *Rand* way to select the examples of the clusters. Due to space limitations, in Figure 1, only the CSS results obtained, when testing with 70% of the datasets, are shown. Despite this, the results for the remaining ratios were very similar.



(a) Ads dataset

(b) Credit dataset

(c) Hapt dataset

(d) Isolet dataset

Figure 1: Results of the 4 datasets with testing set of 70% of the data.

The results showed that the number of clusters, $k$, has a notorious impact on the CSS, mainly on the Ads and Hapt datasets. It is possible to see the improvement of $mk+r$ and $mk$ sampling methods compared to the random sampling, the $r$ method. However, this is not so visible for the Ads and Isolet dataset. In the latter case, this can be due to the fact that the dataset is perfectly balanced, that is, each class has the same number of examples, which allows the random sampling to obtain good results. The Hapt dataset is also well balanced, however, it has a higher number of examples per class. Random sampling seems to be a good choice when dealing with datasets that contains low examples per class. In addition, the best method to select the examples from the clusters showed up to be the Rand.

As the heuristic starts by having very small training sets, $D_v$, the training process is fast. Using this incremental approach, it is easier and faster to find out the CSS of a specific dataset. And by looking at these results, $mk+r$ and $mk$ sampling methods could be a good alternative to the random sampling. As for those two sampling methods, the best shows up to be $mk$ since it presents more consistent results in the first 3 datasets. With these promising results, it is expected that good results will also be obtained for larger datasets.

Despite all this, the reduction for all datasets is noticeable. This shows that most of the data present in these datasets may be redundant, or even irrelevant.

All the datasets present a CSS, and it was possible to discover it by means of a simple heuristic. Therefore, for future work, this heuristic can be a good approach for reducing datasets before using them for data mining tasks.

## 4 Conclusions

In this work we conducted an empirical study on four datasets, using a simple heuristic method, to assess if datasets possess a Critical Sampling Size. Its existence could contribute to data mining tasks, by improving their speed while maintaining the desirable performance.

The experimental results showed the existence of an apparent CSS for each dataset; It is "apparent" since finding the exact CSS is highly intractable, hence whatever sampling generated by the heuristic methods would be an approximated CSS at best. In addition, the CSS of the datasets are demonstrated to be considerably smaller than their initial size. These results seem to validate our assumption that since it is possible to find the (apparent) CFD of a dataset using simple heuristic methods, it is likely that such methods will prove practicable in finding the dataset's (apparent) CSS.

Rather than training a model in the traditional way, by randomly selecting a percentage of the dataset, the proposed heuristic method produced satisfactory results showing that it can significantly reduce the size of the training set, while maintaining the same performance. With these results it proved to be a good alternative to random sampling and a simple and efficient method to discover a CSS for a dataset.

There is always room for improvement. Many aspects of the proposed heuristic deserve further study. Studying more values for the different parameters can provide better conclusions, and the same applies to datasets. Future work can begin by studying more datasets, with even larger dimensions. After that, combining both heuristics to find CFD and CSS, can lead to results that can further reduce datasets.

## References

[1] Ritu Agarwal and Vasant Dhar. Big data, data science, and analytics: The opportunity and challenge for is research, 2014.

[2] D Suryakumar Divya, AH Sung, and Q Liu. The critical dimension problem: No compromise feature selection, eknow 2014. In *Proceedings of the Sixth International Conference on Information, Process, and Knowledge Management*, pages 145–151, 2014.

[3] M. Lichman. Uci machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

[4] Andrew H. Sung, Bernardete Ribeiro, and Qingzhong Liu. Sampling and evaluating the big data for knowledge discovery. In *Proceedings of the International Conference on Internet of Things and Big Data*, pages 378–382, 2016.

# Generating Benchmark Datasets for Intrusion Detection Systems

João Santos[1]
2140141@my.ipleiria.pt

Catarina Silva[1,2]
catarina@ipleiria.pt

Mário Antunes[1,3]
mario.antunes@ipleiria.pt

[1]School of Technology and Management, Polytechnic Institute of Leiria, Portugal

[2]Center for Informatics and Systems of the University of Coimbra, Portugal

[3]Center for Research in Advanced Computing Systems, INESC-TEC, University of Porto, Portugal

## Abstract

Intrusion detection Systems (IDS) are applications used to detect anomalous activities in computer networks and their computers. An emergent challenge faced by the research community is to automate the construction of real-world based datasets with which new detection algorithms can be exploited and benchmarked. The existing solutions to generate datasets are usually based on synthetic network attacks and thus are not representative of real network activity. In this paper, we propose a generic architecture to generate network traffic datasets, including attacks, based on a representation protocol. We also propose a deployment strategy to automate the network traffic datasets construction based on a statistical model of normal behaviour, which allows the generation of diverse and analogous to real world computer networks behaviours.

## 1 Introduction

In network security, an *intrusion* is a set of actions carried on by an *intruder* (internal or external) that attempts to compromise the network infrastructure, through a violation of its security foundations, namely confidentiality, integrity and availability. In general, *intrusion detection* is the process devoted to monitor the network (or systems) events and to trigger an alert to those that may be the target of an ongoing attack.

Intrusion Detection Systems (IDS) are applications usually focused on incidents where events usually occur at very high rates and in an obscured way. Signs of network intrusion can be found by analysing network packets and their associated information flows. Therefore, the main goal of an IDS is to analyse, in real time, network packets in transit, to positively identify all occurrences of actual attacks, and, at the same time, not be mistaken by regular events or be distracted by the signalling of falsely identified attacks.

To accomplish this goal, IDS monitor network traffic in real time and analyse user and system activities, further auditing the faults and vulnerabilities to which the system is exposed. These systems also recognise an activity model and statistically model abnormal behaviour.

Traditionally, IDS deployment has been classified into two distinct detection methods [1]: (i) *signature-based*, as these systems are based on the description of known attacks by the means of a "signature" or a pattern of known and previously seen attacks; and (ii) *behaviour-based*, in which the system builds a model of normal behaviour for the network and then looks for anomalous activities, that is those that do not match the previously established profile.

Snort (www.snort.org) is the most widely used non-commercial open source signature based IDS, being an anomaly based IDS confined mainly to the research community with little expression in production systems [2].

New IDS deployment research needs to be tested against real world network traffic. However, as intruders usually leave slim or no traces of their activity, and real network packets can carry highly sensitive information, it becomes very difficult to have representative data of real attacks, collected from real data networks. To overcome this limitation, artificially created datasets with "safe" and "sanitised" data have been made available to the research community, as DARPA 1999 KDD Cup Challenge dataset [3] and Massicote *et al.* work [4] are two popular and widely used examples.

In this paper, we propose a generic architecture to generate real-world based datasets of network traffic, with both normal and abnormal network traffic flows, that could be used to test, deploy and benchmark intrusion detection systems. The datasets are built upon a statistical model that shapes the distribution of real raw packets flows through time. The dataset can be in raw format or *text based*, depending the user needs.

The adopted strategy tries to overcome some limitations of the existing artificial datasets generators and to expedite the tests of new detection algorithms, using real world based network traffic.

## 2 Generating Datasets

In this section we describe the proposed architecture to generate datasets. We detail the overall architecture, the statistical analysis module, the dataset generation process and the representation protocol used for the resulting datasets.

## 2.1 Proposed architecture

Figure 1 depicts the main blocks of the general architecture we have designed to generate benchmark datasets for IDS testing.

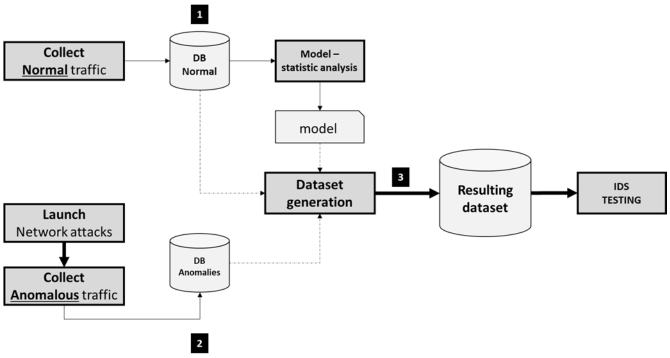


Figure 1: General architecture for data generation

The datasets start to be processed with a normal network traffic collection in a real production network that should have diversity of networking protocols traffic and heterogeneity of network devices.

Regarding the collection of packets related with network attacks, a set of attacks is then launched against a test network composed by, at least, the computers/applications corresponding to the attacker and the victim. In our tests we have launched the attacks with Metasploit Framework (www.metasploit.com), which has a wide range of procedures to exploit already known vulnerabilities.

The collection of packets related with normal and anomalous activities is raw, in the *packet capture* (PCAP) format, and was obtained through the use of specific applications tools, like wireshark (www.wireshark.org) and tcpdump (www.tcpdump.org).

The statistics analysis block (detailed in Section 2.2) produces a profile of the network regarding the amount of network flows and packets processed. The dataset generation block is then fed with the datasets obtained during the network traffic captures, namely normal and anomalous datasets, together with the normal traffic model obtained. According to the input parameters used to model the dataset creation process, namely the frequency of packets flows and its duration, the system will generate the resulting dataset that interleaves the attacks in the normal traffic in a timely order.

The adopted strategy allows the user to freely generate distinct datasets according to the input parameters that shapes the behaviour of the artificial network flows.

## 2.2 Deployment strategy

The first step after capturing the traffic is the analysis and processing of the packets, that will be the base of construction of the profiles related to normal traffic. Such analysis has two phases: (i) aggregation of the flows and (ii) constructing the profiles (see Figure 2). In the first phase flow information is stored in a hash array with source IP; destination IP; protocol; packets.

The statistics analysis block (see Figure 1) analyses the normal network traffic and generates the corresponding profile, namely the amount of packets collected and network flows characterization. Such profile consists of fitting a normal curve.

Figure 2 depicts the statistics module processing. It starts by opening the input file with the network traffic and then processing the corresponding packets. Network traffic profiles are generated according to the network flows, protocols observed and the corresponding amount of packets collected.

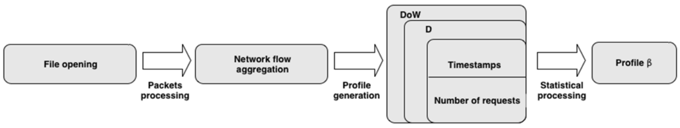


Figure 2 – statistics module processing

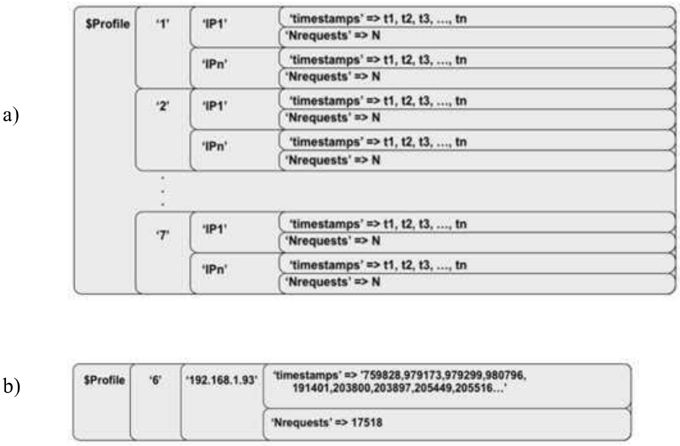
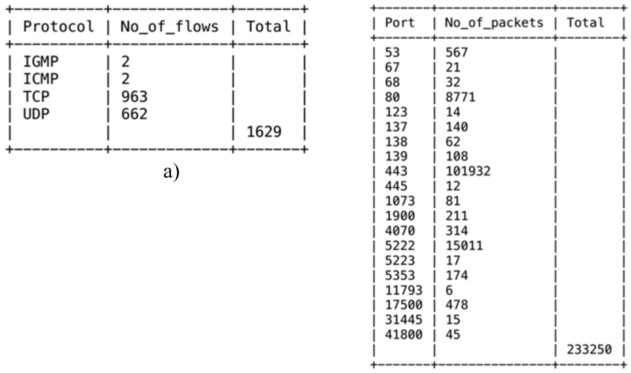

a)
$Profile
'1'
'IP1'
'timestamps' => t1, t2, t3, ..., tn
'Nrequests' => N
'IPn'
'timestamps' => t1, t2, t3, ..., tn
'Nrequests' => N
'2'
'IP1'
'timestamps' => t1, t2, t3, ..., tn
'Nrequests' => N
'IPn'
'timestamps' => t1, t2, t3, ..., tn
'Nrequests' => N
'7'
'IP1'
'timestamps' => t1, t2, t3, ..., tn
'Nrequests' => N
'IPn'
'timestamps' => t1, t2, t3, ..., tn
'Nrequests' => N
b)
$Profile
'6'
'192.168.1.93'
'timestamps' => '759828,979173,979299,980796, 191401,203800,203897,205449,205516...'
'Nrequests' => 17518

Figure 3: Format used to store network activity profiles

Finally, all users (D) with the same day (DoW) are aggregated in a single cumulative distribution function, and with the average of requests the inverse transform of the cumulative distribution function is obtained taking into account the corresponding traffic distribution and using a Weibull distribution as in [5]. The set of these data constitutes the generated profile. The format used to store the profiles is depicted in Figure 3.

Figure 4a) illustrates an example of a *per flow* analysis, identifying the number of flows for each protocol. In Figure 4b) the analysis is made in a *per packet* basis, as it shows the amount of packets processed for each TCP/IP application.



```
+----------+-------------+-------+
| Protocol | No_of_flows | Total |
+----------+-------------+-------+
| IGMP     | 2           |       |
| ICMP     | 2           |       |
| TCP      | 963         |       |
| UDP      | 662         |       |
|          |             | 1629  |
+----------+-------------+-------+
```

a)

```
+-------+---------------+--------+
| Port  | No_of_packets | Total  |
+-------+---------------+--------+
| 53    | 567           |        |
| 67    | 21            |        |
| 68    | 32            |        |
| 80    | 8771          |        |
| 123   | 14            |        |
| 137   | 140           |        |
| 138   | 62            |        |
| 139   | 108           |        |
| 443   | 101932        |        |
| 445   | 12            |        |
| 1073  | 81            |        |
| 1900  | 211           |        |
| 4070  | 314           |        |
| 5222  | 15011         |        |
| 5223  | 17            |        |
| 5353  | 174           |        |
| 11793 | 6             |        |
| 17500 | 478           |        |
| 31445 | 15            |        |
| 41800 | 45            |        |
|       |               | 233250 |
+-------+---------------+--------+
```

b)

Figure 4: Statistics analysis of network traffic (a) by packets (a) and (b) by applications

## 2.3 Dataset generation

The overall dataset generation process, that is network capture, statistics analysis and dataset construction, is carried out through a Perl script. The general algorithm is illustrated in Figure 5.

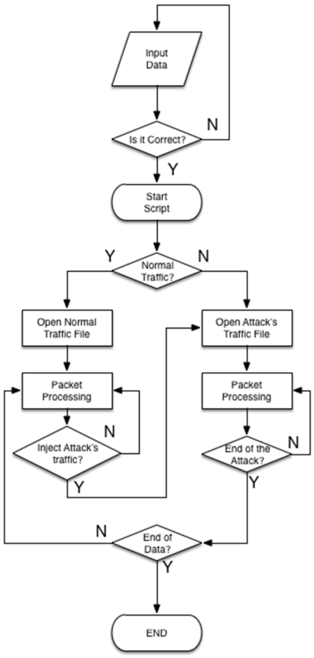


Figure 5: Overall algorithm for datasets construction

The resulting flows can be ready to be tested in real network environments, with the purpose of determining the capabilities of IDS in an organization or can be used as a research tool to construct benchmarks for new algorithm testing. In the later case, a text-based protocol is defined, as the user is able to dump to a text file the contents of each TCP/IP header fields, according to the input data needed to test a specific IDS. The fields chosen to integrate the dataset should be written in a stream of characters, without spaces and separated by a comma. In this preliminary stage, the fields available are the following: source IP address (`sip`), destination IP address (`dip`), source port (`sp`), destination port (`dp`), packet payload and timestamp (`ts`).

## 3 Conclusions and Future Work

In this work we have focused on an architecture to generate benchmarks for intrusion detection systems. The proposed approach is able to generate real-world based datasets of network traffic, with both normal and abnormal network traffic flows that statically represent real flows in an organization. Such datasets can be used as benchmark to test, deploy and benchmark intrusion detection systems. The datasets are built upon a statistical model that shapes the distribution of real raw packets flows through time. Moreover, the architecture allows for outputs in raw format to directly test working intrusion detection systems or can be text-based, allowing further research in innovative algorithms for intrusion detection.

Future work is foreseen in further testing the architecture, both in real intrusion scenarios, as well as in developing new detection algorithms.

## References

[1] Liao, H. J., Lin, C. H. R., Lin, Y. C., & Tung, K. Y., “Intrusion detection system: A comprehensive review”, Journal of Network and Computer Applications, 36(1), 16-24, 2013.

[2] Wu, S. X., & Banzhaf, W., “The use of computational intelligence in intrusion detection systems: A review”, Applied Soft Computing, 10(1), 1-35, 2010.

[3] Lippmann, R., Haines, J. W., Fried, D. J., Korba, J., & Das, K., “The 1999 DARPA off-line intrusion detection evaluation”, Computer networks, 34(4), 579-595, 2000.

[4] Massicotte, Frederic, et al. “Automatic evaluation of intrusion detection systems”, Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual. IEEE, 2006.

[5] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, “Toward developing a systematic approach to generate benchmark datasets for intrusion detection”, Comput. Secur., vol. 31, no. 3, pp. 357–374, 2012.

# User Specific Adaptation in Automatic Transcription of Vocalised Percussion

António Ramires[1][2]
antonio.ramires@inesctec.pt
Rui Penha[1][2]
rui.penha@inesctec.pt
Matthew E. P. Davies[1]
mdavies@inesctec.pt

[1] INESC TEC
Sound and Music Computing Group
Rua Dr. Roberto Frias, s/n, 4200 - 465 Porto, Portugal

[2] Faculty of Engineering
University of Porto
Rua Dr. Roberto Frias, s/n, 4200 - 465 Porto, Portugal

## Abstract

The goal of this work is to develop an application that enables music producers to use their voice to create drum patterns when composing in Digital Audio Workstations (DAWs). An easy-to-use and user-oriented system capable of automatically transcribing vocalisations of percussion sounds, called LVT - Live Vocalised Transcription, is presented.[1] LVT is developed as a Max for Live device which follows the "segment-and-classify" methodology for drum transcription, and includes three modules: i) an onset detector to segment events in time; ii) a module that extracts relevant features from the audio content; and iii) a machine-learning component that implements the k-Nearest Neighbours (kNN) algorithm for the classification of vocalised drum timbres.

Due to the wide differences in vocalisations from distinct users for the same drum sound, a user-specific approach to vocalised transcription is proposed. In this perspective, a given end-user trains the algorithm with their own vocalisations for each drum sound before inputting their desired pattern into the DAW. The user adaption is achieved via a new Max external which implements Sequential Forward Selection (SFS) for choosing the most relevant features for a given set of input drum sounds.

The evaluation of LVT addresses two objectives. First, to investigate the improvement in performance with user-specific training, and second, to assess if LVT can provide an optimised workflow for music production in Ableton Live when compared to existing drum transcription algorithms. Obtained results demonstrate that both objectives are met.

## 1 Introduction

The development of computers' performance capacity, and the consequent possibility for real-time Digital Signal Processing (DSP) for audio, led to the appearance of Digital Audio Workstations (DAWs), making the creation of computer music available to the general public. Following these advances, many new instruments and interfaces for creating electronic music have surfaced. With changes in music culture, music production and how musicians work with their instruments has also changed. In other words, the ability to invent and reinvent the way to produce music is key to progress. Consequently, new proposals are necessary, such as designing new techniques for the composition of music.

Within the genre of Electronic Music, sequencing drum patterns plays a critical role. However, inputting drum patterns into DAWs often requires high technical skill on the part of the user, either by physically performing the patterns by tapping them on MIDI drum pads, or manually entering events via music editing software. For non-expert users both options can be very challenging, and can thus provide a barrier to entry. However, the voice is an important and powerful instrument of rhythm production, so it can be used to express or "perform" drum patterns in a very intuitive way - so called "beatboxing." In order to leverage this concept within a computational system, our goal is towards a system to help users (both expert musicians and amateur enthusiasts) input rhythm patterns they have in mind into a sequencer via the automatic transcription of vocalised percussion. Our proposed tool is beneficial both from the perspective of workflow optimisation (by providing accurate real-time transcriptions), but also as means to encourage users to engage with technology in the pursuit of creative activities. From a technical standpoint, we seek to build on the state of the techniques from the domain of music information retrieval (MIR) for drum transcription [2, 4] but actively targeted towards end-users and real-world music content production scenarios.

## 2 Methodology

A vocalised drum transcription software, LVT, able to be trained with the user vocalisations is proposed. LVT is developed as a Max for Live project – a visual programming environment, based on Max 7[2], which allows users to build instruments and effects for use within the Ableton Live[3] DAW. To develop LVT, a dataset of vocalised percussion was compiled. A group of 20 participants (11 male, 9 female) were asked to record two short vocalised percussion tracks, one identical for all participants, and the other, an improvised pattern. These input percussion tracks were recorded three times: on a low quality laptop microphone, on an iPad microphone, and using a studio quality microphone (AKG c4000b). All recorded audio tracks were manually annotated using Sonic Visualiser[4], a free application for viewing and analysing the contents of music audio files. The participants spanned a wide range of experience in beatboxing (from beatboxing experts, to those who had never vocalised drum patterns before), and covered a wide age range. Thus, we consider the annotated dataset to be representative of a wide range of potential users of the system, and highly heterogeneous in terms of the types of drum sounds.

Our proposed vocalised percussion transcription system was developed following a user-specific approach. LVT follows the "segment and classify" method for drum transcription [2] and integrates three main elements: i) an onset detector – to identify when each drum sound occurs, ii) a component that extracts features for each event, and iii) a machine learning component to classify the drum sounds. In the Max for Live environment, the onset detection was performed with `AubioOnset∼`[5]. Feature extraction was performed in real-time using existing Max objects: `Zsa.mfcc∼` – to characterise the timbre, `Zsa.descriptors` [3] – to provide spectral centroid, spread, slope, decrease and rolloff features [3], and finally the zero crossing rate and number of zero crossings were calculated with the `zerox∼` object. The machine learning component is trained with the user's preferred vocalisation and the features are selected which give the best results for the provided input. This is achieved using the Sequential Forward Selection method [5] along with a k-Nearest Neighbours classification algorithm, with the most significant features selected by the accuracy obtained from testing the training data (in our case, the annotated improvised patterns from each participant). SFS works by selecting the most significant feature, according to a specific parameter (in this case the classification accuracy), and adding it to an initially empty set until there are no improvements or no features remain. The k-NN algorithm was implemented using timbreID[1], and a new external for Max was developed to implement the SFS. A user interface was created in Max for Live to facilitate the utilisation of the application by end-users. A screenshot of the interface of LVT is shown in Fig. 1. It demonstrates the user-specific training stage – where a user inputs a set number of the drum timbres they intend to use, after which their vocalised percussion is transcribed and rendered as a MIDI file for subsequent synthesis.

To operate LVT, a user loads the device in Ableton Live and then vocalises the set of desired drum sounds they intend to use, e.g. five kick sounds followed by five snare sounds, followed by five hi-hat sounds. Once the expected number of drum sounds have been detected, the SFS algorithm then identifies the subset of features which best separate the drum sounds for the user. After training, the user can then vocalise rhythmic patterns which are automatically converted from audio to a MIDI representation in the DAW for later synthesis and editing.

[1] This work is derived from the MSc dissertation of António Ramires, conducted in the Department of Electrical and Computer Engineering in the Faculty of Engineering, University of Porto.

[2] www.cycling74.com

[3] http://www.ableton.com/en/

[4] http://www.sonicvisualiser.org/

[5] https://aubio.org/manpages/latest/aubioonset.1.html

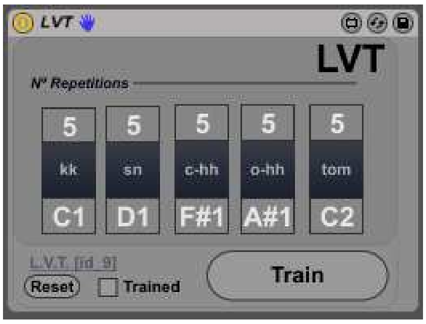


Figure 1: User interface of the LVT device.

Table 1: Number of operations and F-measure for the AKG microphone.

| | *Edit Operations* | | | *F-measure* | | |
|---|---|---|---|---|---|---|
| | Modify | Add | Remove | Kick | Snare | Hi-hat |
| Ableton | 33 | 12 | 296 | 0.518 | 0.470 | 0.297 |
| LDT | 52 | 24 | 206 | 0.538 | 0.204 | 0.419 |
| LVT | 39 | 7 | 15 | 0.914 | 0.691 | 0.802 |

## 3 Results

The evaluation of LVT was designed to serve two purposes. First, to understand how a user-specific trained system performs against state of the art drum transcription system (which have been optimised over large datasets without any user-specific training), and second, to explore how LVT could improve a producer's workflow. We compared LVT against two existing drum transcription algorithms: LDT [4], and Ableton Live's built-in "Convert Drums to MIDI" function. For validation data we used the non-improvised vocalised patterns from our annotated dataset.

To compare the accuracy of the systems we use the F-measure of the transcriptions. Then, to investigate how our system could improve a producers workflow, the "effort" to get an accurate transcription was calculated by counting the number of editing operations required to obtain the desired patterns. These operations are as follows: to *modify*, to *add*, or to *remove* a MIDI note.

Table 1 summarises the results obtained from counting the total number of operations needed to obtain the desired pattern for the testing data recorded on the studio quality AKG c4000b microphone and the corresponding F-measure per vocalised drum sound, on the three drum transcription systems. The results demonstrate that, for the studio quality microphone, vocalised drum transcription accuracy for LVT is substantially higher than the other systems, and far fewer modifications were required to obtain the desired patterns when editing the automatic transcriptions.

To see the effect of user-specific training on the performance of LVT, an example is provided where LVT is trained on one user and tested on another – and vice-versa. When training the LVT with a different person with different vocalisations, the accuracy of the transcription is decreased as shown in Fig. 2. In the upper part of each screenshot is the transcription of the user when trained with its own vocalisations, while the bottom part corresponds to the transcription when trained with the other user. As can be seen, without the user-specific training, many misclassifications occur.

By examining the previously obtained results, we infer that LVT can provide a transcription closer to the ground truth than the existing state of the art systems, as shown by the higher F-measure. In addition to LVT being trained per individual user, these results may also derive from the fact that LVT does not try to detect polyphonic events (more than one drum vocalisation at the same time) as the other systems do. Furthermore, LVT does not detect as many events as the other systems, and this has a strong influence on the number of false positives, and hence the F-measure. The number of events to achieve the desired transcription, presented in Table 1, shows that the end-user of the system does not have to perform as many actions when producing music, which has a positive impact on the workflow, leaving more time for creative experimentation.

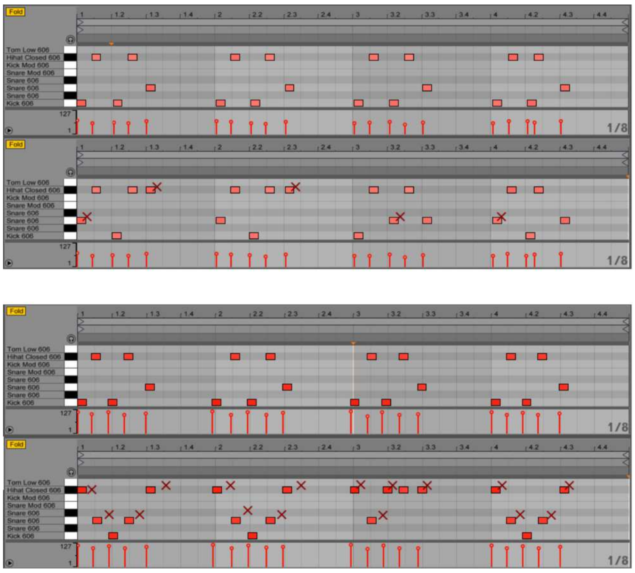


Figure 2: (top) First user vocalisations trained with the second user. (bottom) Second user vocalisations trained with the first user.

## 4 Conclusions

In this paper, we have presented LVT – a new interface for assistive music content creation. LVT allows Ableton Live users to sequence MIDI patterns that can be used for designing and performing rhythms with their voice. Existing state of the art systems, including one already in Ableton Live, are not able to transcribe vocalised percussion as effectively because these tools are trained for general recorded drum sounds which are typically not vocalised. Indeed, because different people vocalise drum sounds in different ways, LVT explicitly seeks to model and capture this behaviour via user-specific training. Our evaluation shows LVT to be very effective for wide range of users and vocalisations, outperforming existing systems. Furthermore, we believe LVT can be applied to any kinds of arbitrary non-pitched percussive sounds – provided that the training sound types are sufficiently different from one another, and can thus be well separated in the audio feature space using SFS.

LVT is implemented as a Max for Live device, and thus fully integrates into Ableton Live, allowing users of all ability ranges to experiment with music sequencing driven by their own personal percussion vocalisations within an easy-to-use graphical user interface.

## 5 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-01-0145-FEDER-006961», and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013.

Project TEC4Growth-Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01-0145-FEDER-000020 is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

## References

[1] W. Brent. A timbre analysis and classification toolkit for pure data. In *Proc. of ICMC*, pages 224–229, 2010.

[2] O. Gillet and G. Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3):529–540, March 2008.

[3] M. Malt and E. Jourdan. Zsa. descriptors: a library for real-time descriptors analysis. In *Proc. of 5th SMC Conference*, pages 134–137, 2008.

[4] M. Miron, M. E. P. Davies, and F. Gouyon. An open-source drum transcription system for Pure Data and Max MSP. In *Proc. of ICASSP*, pages 221–225, May 2013.

[5] A. W. Whitney. A direct method of nonparametric measurement selection. *IEEE Trans. Comput.*, 20(9):1100–1103, September 1971.

# An audio-only method for advertisement detection in broadcast television content

António Ramires
antonio.ramires@inesctec.pt
Diogo Cocharro
diogo.m.cocharro@inesctec.pt
Matthew E. P. Davies
mdavies@inesctec.pt

INESC TEC
Sound and Music Computing Group
Rua Dr. Roberto Frias, s/n
4200 - 465 Porto, Portugal

## Abstract

We address the task of advertisement detection in broadcast television content. While typically approached from a video-only or audio-visual perspective, we present an audio-only method. Our approach centres on the detection of short silences which exist at the boundaries between programming and advertising, as well as between the advertisements themselves. To identify advertising regions we first locate all points within the broadcast content with very low signal energy. Next, we use a multiple linear regression model to reject non-boundary silences based on features extracted from the local context immediately surrounding the silence. Finally, we determine the advertising regions based on the long-term grouping of detected boundary silences. When evaluated over a 26 hour annotated database covering national and commercial Portuguese television channels we obtain a Matthews correlation coefficient in excess of 0.87 and outperform a freely available audio-visual approach.

## 1 Introduction

The classification of audiovisual content into categories and the identification of advertising has become increasingly important for end-users, broadcasters and entities that have contracted advertising space. This has special importance in the case of television content both for the need to archive content with the advertising removed, and in streaming contexts to allow for the region-specific substitution of advertising.

Currently, the delimitation of advertising segments i.e., the identification of the beginning and end moments of a contiguous set of advertising content is typically performed by a human operator. As a result, the process is labour-intensive, expensive, and potentially error-prone [2]. One means to improve the workflow of the human operator is to provide an automatic analysis of the broadcasting content which can classify the time-line into regions of advertising and regular programming.

Existing algorithms for the detection of advertising fall into two main categories. Those which use explicit prior knowledge of a known set of advertisements and identify them using fingerprinting methods [1], and those which rely on heuristics as advertising indicators. For both types of approach, information can be leveraged from the video signal alone (logos, black frames, scene changes etc.), or in combination with the audio stream within audio-visual approaches [3].

In this work, our focus is on audio-only approach for television advertising detection which makes no use of video information, meta-data concerning the content type, or any prior knowledge of which advertisements can appear. To this end, we seek to discover if there is sufficient information in the audio signal alone to locate where advertisements occur. From this perspective, relevant acoustic cues include the presence of silence (often co-occurring with black frames at content boundaries), the presence of jingles, fast paced narration, background music, and identified repeated content – which operates on the assumption that advertisements are repeated more frequently than regular programming.

In our audio-based approach, we focus on a single acoustic property, that of silence – which we assume to indicate very low signal energy rather than digital zeros in the audio bit-stream. We believe that silences have been under-used in the existing literature having been treated as just one feature among many which contribute towards the final decision. In our approach, we seek to maximize the information that can be obtained from detecting silences. Furthermore, we propose that by effective characterisation of different types of silences and the large scale grouping of an identified set of "boundary silences" we can obtain a very reliable descriptor of advertising boundaries in television.

## 2 Approach

Our approach centres on the existence and detection of short pauses of silence (i.e., very low audio signal energy) in between separate pieces of content. We now provide an overview of each stage of the algorithm. Throughout, we assume the audio signal (a stereo signal sampled at 48kHz with 24-bit precision) has already been separated from the video content, and mixed down to mono. We notate the audio input as, $x$.

To maintain parity with video frame rate of the television content (and allow easy integration with future video-based analysis) we partition $x$ into non-overlapping audio frames of 1920 samples (equivalent to 25 video frames per second). In each audio frame, $x_i$, we calculate the signal energy, $e_i = 20\log_{10}\left(\sqrt{\text{mean}(x_i{}^2)}\right)$. By taking the measurement in dB, we force all low energy parts of the signal to take large negative values. Next, to find all the low energy points in the input signal, we compare $e_i$ at each frame, $i$ to a silence threshold, $\eta$=-60 dB, and retain those frames $i_s$ for which $e_i \leq \eta$. An example is shown in the top plot of Fig. 1.

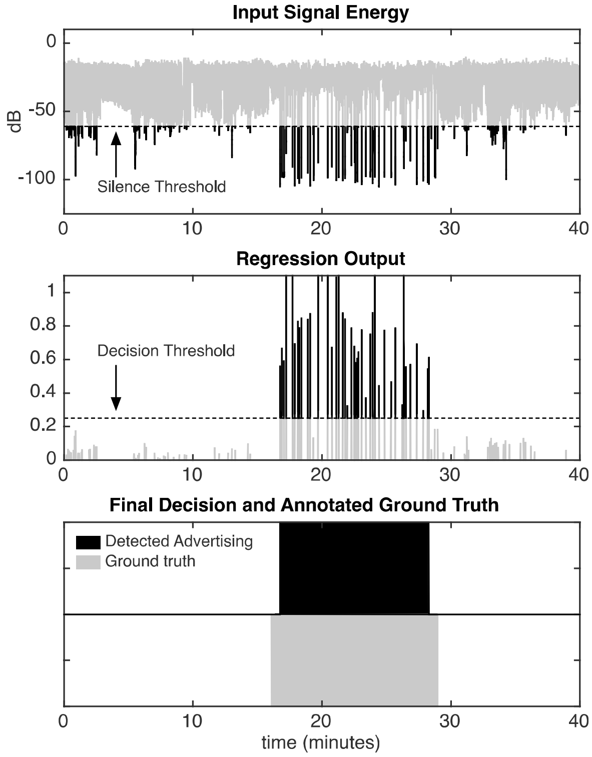


Figure 1: (top) Energy of input signal, with regions under the silence threshold shown in black. (middle) Output regression model on detected silences. Points above the decision threshold are shown in black. (bottom) The output classified as advertising and the corresponding ground truth.

Since short regions of silence can occur naturally within programming, e.g. as pauses between speech (either during narration or interviews with no background music or noise), we must filter out those silences which do not correspond to content boundaries. In our model we assume a boundary silence to be: short in duration, have a low minimum value, and be surrounded by regions of much higher energy. From a broadcast perspective we understand this is perceptually loud advertising content either side of a brief, imperceptible drop in energy, as shown in Fig. 2.

To distinguish between different types of silence we collect a small set of statistics: the max, mean, min, inter-quartile range, standard deviation, skewness, and kurtosis from a small temporal window of ±6 s (± 150

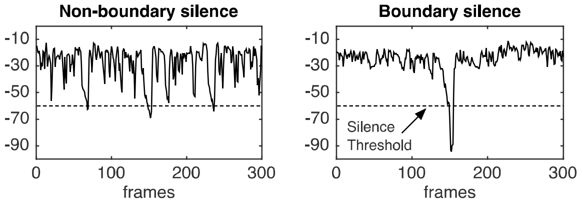


Figure 2: Examples of non-boundary silence (left) and boundary silence (right). The detected silence is at the mid-point of each plot.

audio frames) of the energy signal $e$ surrounding each detected silence $i_s$. We then perform a basic multiple linear regression on the extracted features where positive examples (i.e. annotated boundary silences) are labelled as 1, and non-boundary silences are labelled as 0. The output of the regression is shown in the middle plot of Fig. 1. Here, all detected silences greater than the decision threshold, $\beta$=0.25, are retained and set to a value of 1, with all others discarded.

In the final stage of our algorithm, we pass a sliding rectangular window of 150 s duration across the thresholded regression output. We determine regions of advertising as those which adhere to the following two conditions: i) there is more than one detected boundary silence within the long-term window (i.e. at least one starting and one ending silence); ii) the total duration of any period of advertising must be at least 60 s. In this way isolated silences or those which are far from one another are excluded. The start of the detected advertising region is marked at the frame where the first detected boundary silence exits the long-term window. Likewise the end of the region occurs at the frame when the final boundary silence of any group exits the long-term window. An example of final output of the system is shown in the bottom plot of Fig. 1.

## 3 Results and Discussion

We evaluate our algorithm over an annotated dataset we have compiled covering national (two instances of RTP 1 and one of RTP 2) and commercial channels (SIC and TVI) of Portuguese television. The dataset contains over 26 hours of content (segmented in 28 programmes), which has been annotated at two levels. First, to mark the high level boundaries between regular programming and advertising blocks, and second at a finer temporal level to marks the boundaries between all commercials. We use this second level for training the linear regression model.

In order to measure the performance of silence-based method, we first count the number of true positives, $T_P$, true negatives, $T_N$, false positives, $F_P$, and false negatives, $F_N$, where a $T_P$ corresponds to a region which is both annotated and detected as advertising.

As we can expect with broadcast television content, a far greater proportion of the content corresponds to scheduled programming rather than advertising (in our case, approximately 12% is advertising). While many approaches in the literature report the F-measure as a performance indicator for advertising, this excludes any information about the number of $T_N$. To incorporate this information we instead report Matthews correlation coefficient, $M$, which is calculated as follows:

$$M = \frac{T_P \times T_N - F_P \times F_N}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}} \qquad (1)$$

In addition to reporting the performance of our proposed approach, we also ran an open source audio-visual approach called ComSkip[1] under the default parameter settings. A comparison of performance between the two approaches is shown in Table 1.

Table 1: Summary of dataset and comparison of algorithm performance.

| Input Channel | Total Duration | Advertising Duration | ComSkip Accuracy | Proposed Alg. Accuracy |
|---|---|---|---|---|
| RTP $1_a$ | 6h52m | 0h23m | 0.426 | 0.782 |
| RTP $1_b$ | 1h10m | 0h11m | 0.007 | 0.863 |
| RTP 2 | 8h25m | 0h24m | 0.366 | 0.499 |
| SIC | 8h37m | 2h18m | 0.648 | 0.931 |
| TVI | 1h13m | 0h9m | 0.794 | 0.966 |
| Overall | 26h17m | 3h24m | 0.610 | **0.874** |

As can be seen, our proposed approach outperforms ComSkip across all channels, with a correlation coefficient in excess of 0.87. Indeed, our approach performs especially well on the commercial channels (SIC and TVI), which contain large blocks of advertising content (running into several minutes at a time) with explicit use of silences between individual advertisements.

The lowest performance was obtained on RTP 2. This channel contained a far lower proportion of commercial advertising, with the breaks between programming more frequently containing trailers for upcoming in-channel content (and without such prominent silence boundaries). Since this content falls between the main programming, it can be understood as advertising, and thus something which our current approach cannot readily detect. However, given the critical requirement in advertising removal applications not to misclassify programming as advertising, our proposed approach has explicitly been parameterised to minimise false positives. To this end, it provides "conservative" estimates of advertising boundaries. Indeed, over the 26 hours, our approach has just 6 false positive frames, with ComSkip having only 761 false positive frames ($\sim$ 30 s).

A potential criticism of the comparative results is that they may be somewhat optimistic since our approach has partial access to the dataset for training, where as ComSkip does not. However, our multiple linear regression model was trained using leave one out cross fold validation at the programme level, and therefore we maintain some separation between training and testing material. Informal tests on currently un-annotated validation data also indicates highly promising performance and larger-scale evaluation will be among the main areas of future work.

## 4 Conclusions

We have a presented a new audio-only approach for the detection of advertising in television broadcast content. Our approaches relies on the short, medium, and long-term modelling of silences within the audio stream as a means for distinguishing regular programming from advertising. A novel feature of our approach is the ability to reject silences (e.g. pauses in speech) which do not exhibit the statistical properties of content boundaries. Currently our approach has been optimised for Portuguese television content, therefore main focus of our future work will be to investigate the accuracy of our approach on international television content. Furthermore, we intend to enhance our audio-only model via the inclusion of other important cues includes in jingle detection, music/speech separation and audio production effects related to bandwidth and dynamic range.

## 5 Acknowledgements

This article is a result of the project MOG CLOUD SETUP - N$^{\text{o}}$17561, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

Diogo Cocharro was supported by Project TEC4Growth-Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01-0145-FEDER-000020 is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

## References

[1] P. Cardinal, V. Gupta, and G. Boulianne. Content-based advertisement detection. In *INTERSPEECH*, pages 2214–2217, 2010.

[2] D. Conejero and X. Anguera. TV advertisements detection and clustering based on acoustic information. In *Intl. Conf. on Computational Intelligence for Modelling Control Automation*, pages 452–457, 2008.

[3] M. Covell, S. Baluja, and M. Fink. Advertisement detection and replacement using acoustic and visual repetition. In *IEEE Workshop on Multimedia Signal Processing*, pages 461–466, 2006.

[1] http://www.kaashoek.com/comskip/, v. 0.82, accessed 06-15-2017.

# Evolutionary Machine Learning: An Essay on Benchmarking

Filipe Assunção
fga@dei.uc.pt
Nuno Lourenço
naml@dei.uc.pt
Bernardete Ribeiro
bribeiro@dei.uc.pt
Penousal Machado
machado@dei.uc.pt

CISUC
Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal

## Abstract

The search for adequate structures and parameters for Machine Learning (ML) models is problem specific and time consuming. Often, researchers follow an iterative trial-and-error process, where suitable values for multiple parameters are tested. One way to address this issue is the application of Evolutionary Computation (EC) to search, optimise and tune the ML models. Selecting appropriate benchmarks for comparing different approaches is not always trivial and is a common problem of both EC and ML. However, when combining both fields it is possible to use the evolutionary process to our advantage, speeding up the evaluation stage. In this paper we discuss what can be done to mitigate some of the issues of benchmarking in Evolutionary Machine Learning (EML). The positions herein presented denote the point of view of the authors and should not be seen as a strict methodology, but rather as a set of guidelines.

## 1 Introduction

Evolutionary Machine Learning (EML) is a sub-field of Artificial Intelligence that applies Evolutionary Computation (EC) to automatically search for the structure and/or parameterisation of Machine Learning (ML) models. Examples of EML works are the evolution of the topology and weights of Artificial Neural Networks (ANNs) [3], or the evolution of multiple Support Vector Machine (SVM) parameters [4]. By adopting the principles behind EC, a population of solutions scattered in the domain space of the ML model is evolved, making it less likely to become trapped in a local optima, which would likely happen if the model was tuned by manually trying different parameter configurations in an attempt to reach near-optimal performances.

When working with EML one of the main challenges concerns benchmarking. In brief words, benchmarking can be defined as the act of performing experiments with datasets to compare the performance of different approaches. From this definition it is clear that the selection of datasets, experimental design and analysis are principles that are linked. In the current essay we will focus primarily on tackling the issues related with benchmarking from the perspective of the datasets, i.e., we will discuss the questions that must be faced when planning the experiments to test new / existing approaches regarding the choice that has to be made regarding the appropriate benchmarks for assessing the quality of the methods, and how many datasets should be used.

If we decide on simple benchmarks, with a low number of features and instances, the methodology will likely find solutions that perform well. But, these results have questionable importance, since solutions for such problems might be easily hand-crafted. On the other hand, when dealing with real world complex problems, mapping the candidate solutions to a comprehensible model, and assessing their quality on such huge datasets can be time consuming, making it impossible to timely measure the quality of the models. In addition, to tackle such problems, we often require large amounts of computational power, making the use of the benchmark unfeasible. Discussion around the problematic of benchmarking is not new. McDermott et al. [7] have already pointed the selection of benchmarks for evaluating Genetic Programming (GP) methods to be one of the open issues in the field.

In this position paper, more than proposing a methodology that should be strictly followed, we aim at discussing good practises. In Section 2 we focus on dataset selection. Next, in Section 3, we investigate methods that try to cope with the challenges posed by big data. To end, in Section 4, open questions and future discussion are raised.

## 2 Datasets

In EML the datasets are commonly grouped according to the ML task that the models being evolved aim to solve: (i) clustering; (ii) regression; or (iii) classification. In addition, and following the structure proposed by Prechelt [9] these benchmarks belong to one of the following categories:

**Artificial** – data is artificially generated following a given equation (logic or arithmetic);

**Realistic** – although the data is also artificially generated (as in the above category) it simulates the rules and specifications of real world systems (e.g., physical models);

**Real problems** – data gathered directly from observing the real world.

More and more contributions to the field have focused on the use of real world problems, which is motivated by the desire to search for true Artificial Intelligence, i.e., systems capable of outperforming the human performance and automate common everyday tasks.

With the increase of computational power and performance of the evolved models, popularised by Graphics Processing Units (GPUs) and consequent emergence of deep learning techniques, the problems that practitioners try to solve are becoming increasingly more challenging. However, there are no well-established methodologies specifying how to select which problems to test on. Although there are works in the literature describing how to measure the complexity of datasets (e.g. [11]), they tend to be difficult, and time consuming to use. It is impractical to apply such methods to a wide range of benchmarks, and thus, authors often base their decision on the complexity in terms of number of instances and dimensionality of the input space / number of features, and on the benchmarks used by the methodologies to which comparisons are going to be established. Another criteria that is often analysed is the available progression margin, which defines the problem complexity based on the difficulty that previous approaches had to solve it.

There are several platforms that work as repositories for benchmarks. The most popular is the UCI ML repository [6], which at the time of writing is composed by 394 datasets. The UCI platform stores information on each dataset, which includes the number of instances, features, types of the data and the task to be performed. In addition, a brief description of each benchmark and corresponding reference papers are provided. Although it provides a good platform in the sense that it allows users to scan a large list of benchmarks, showing their main characteristics, it does not provide a list of the results obtained by previous methodologies in a clean and accessible way. Furthermore, despite the large number of available benchmarks, a large percentage (approximately 40%) has less than 1000 instances, and about 23% have no more than 10 features. A platform that solves one of the issues found in the UCI repository is Kaggle (check `http://www.kaggle.com`): a web-platform for the organisation of contests often tackling real world problems. By using leaderboards the performance of different approaches in each benchmark becomes clear. The benefits of the the two previous platforms are combined in OpenAI [2]; the main disadvantage of OpenAI is that it is focused on reinforcement learning problems, more particularly, game environments.

From the above discussion on the available platforms we suggest that the ideal platform should at least follow the following principles:

- Provide a detailed description of each dataset, including properties such as the number of features and instances, task to be performed, and type of the dataset. Complexity of the dataset according to established metrics should also be provided;

- The performances obtained on each benchmark by different approaches should be shown and detailed in the form of a list. Each result entry should be accompanied by the article describing the method and whenever possible the implementation. Authors should be allowed to submit this information so that the platform is self-maintaining;

- Ideally, the platform should provide means to confirm the accuracy of the results by automatically running new experiments with the provided code on different partitions of the benchmark;

- It should be possible to order and filter the benchmarks available in the platform according to any of its properties and results, so that one can easily explore them and choose the ones to tackle.

So far we have discussed the main challenges in the analysis and decision of which benchmarks to use for testing purposes. But, the question of how many benchmarks should be used has not yet been addressed. An obvious answer would be that the more datasets are used the better, so that it is easier to characterise the behaviour of the tested method on benchmarks with different properties. However, this might not be feasible: papers have limited sizes, and the time needed for conducting such an amount of experiments makes authors inclined to select a small amount of benchmarks. Specially if we are dealing with real world problems, where the available amounts of information often comprise Big Data (further discussed in the next section). We recommend that experiments should be conducted in at least four benchmarks (the more the better). Testing on fewer than that does not allow for any strong conclusions about the quality of the approach rather than the one that it performs better or worse in a couple datasets, but an extrapolation and generalisation assumption to other benchmarks can hardly be made. Moreover, it is our opinion that the benchmarks should be selected with an increasing complexity degree, so that it is possible to test if the approach despite performing good in difficult problems also leads to good results in simple and easier tasks, and vice-versa.

## 3 Dealing with Big Data

By combining the principles of EC with ML, a population of candidate solutions encoding the model's structure and/or parameters is evolved through time. Although evolution is paralallelisable, assessing the quality of each candidate solution in real world problems can be time consuming.

Several tools that take the advantages provided by GPU computing have been proposed recently (e.g., Caffe [5] or Tensorflow [1]). Using these frameworks has two main advantages: in the one hand, they provide stable implementations for evaluating the performance of the evolved models; on the other hand, by providing GPU interfaces they speed up the evaluation time.

Nevertheless, in some circumstances the speed up introduced by the use of GPU processing is not enough. Imagine evaluating a deep neural network that takes about 1 hour to train; if a population of 100 candidate solutions is evolved then each generation would take about 4 days, which makes evolution unfeasible. Thus, when this happens authors normally resort to sampling techniques, and smaller training sessions that give some insight on the expected performance of the model on the long term. We believe that random sampling approaches are not the most appropriate form to reduce the dimension of datasets, as they may fail to retain some of the properties of the benchmark, possibly leading to deceiving results. We defend that we should use the evolutionary process in our favour, and sample a percentage of the instances of the benchmark every given number of generations, taking the results on the samples into account when generating new ones. Examples of these type of approaches have been proposed by Stanovov et al. [10] and Morse and Stanley [8].

## 4 Road Ahead

In this short essay we have pointed out various aspects of current benchmarking practices in EML. We have discussed the following issues:

- Limitations imposed by the difficulty of selecting a set of benchmarks for testing a developed (or existing) methodology. It is not clear what makes a dataset complex; most practitioners make such decisions based on benchmark properties, such as number of instances and dimensionality of the input space, or based on the performance of other approaches;

- Similar to the selection of benchmarks, the same rationale can be applied to deciding how many datasets should be used for conducting experiments. We defend that at least four different datasets, with different complexities should be used.

- Dealing with Big Data is challenging, specially when multiple candidate solutions are being evolved in simultaneous, and need to be evaluated for assessing their performance. To deal with this limitation the dataset can be sampled, but taking advantage of the iterative nature of EC.

Nonetheless, there are many more questions that have to be addressed by further research in the area, which directly impact how the evolved models are selected and compared. One of the most important ones comprises the definition of metrics that can objectively measure the difficulty of benchmarks. We are well aware that this is not an extensive review, and that there are several works that already follow some of the guidelines discussed here. Our main goal with this essay is to set off a discussion about good practices that can improve the field, leading to better and sound research.

## Acknowledgements

This work is partially funded by: Fundação para a Ciência e Tecnologia (FCT), Portugal, under the grant SFRH/BD/114865/2016.

## References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI gym. *arXiv preprint arXiv:1606.01540*, 2016.

[3] Dario Floreano, Peter Dürr, and Claudio Mattiussi. Neuroevolution: from architectures to learning. *Evolutionary Intelligence*, 1(1):47–62, 2008.

[4] Frauke Friedrichs and Christian Igel. Evolutionary tuning of multiple SVM parameters. *Neurocomputing*, 64:107–117, 2005.

[5] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[6] M. Lichman. UCI ml repository, 2013. URL `http://archive.ics.uci.edu/ml`.

[7] James McDermott, David R White, Sean Luke, Luca Manzoni, Mauro Castelli, Leonardo Vanneschi, Wojciech Jaskowski, Krzysztof Krawiec, Robin Harper, Kenneth De Jong, and Una-May O'Reilly. Genetic programming needs better benchmarks. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation*, pages 791–798. ACM, 2012.

[8] Gregory Morse and Kenneth O. Stanley. Simple evolutionary optimization can rival stochastic gradient descent in neural networks. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference*, pages 477–484. ACM, 2016.

[9] Lutz Prechelt. A quantitative study of experimental evaluations of neural network learning algorithms: Current research practice. *Neural Networks*, 9(3):457–462, 1996.

[10] Vladimir Stanovov, Eugene Semenkin, and Olga Semenkina. Instance selection approach for self-configuring hybrid fuzzy evolutionary algorithm for imbalanced datasets. In *International Conference in Swarm Intelligence*, pages 451–459. Springer, 2015.

[11] Julian Zubek and Dariusz M Plewczynski. Complexity curve: a graphical measure of data complexity and classifier performance. *PeerJ Computer Science*, 2:76, 2016.

# Evolutionary Machine Learning: An Essay on Experimental Design

Filipe Assunção
fga@dei.uc.pt
Nuno Lourenço
naml@dei.uc.pt
Bernardete Ribeiro
bribeiro@dei.uc.pt
Penousal Machado
machado@dei.uc.pt

CISUC
Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal

## Abstract

Evolutionary Machine Learning (EML) combines Evolutionary Computation (EC) with Machine Learning (ML) to automatically search for the best structure and/or parameterisation of ML models for solving specific tasks. However the results reported by the authors in their articles detail their work, replicating the results and comparing them to other approaches are tasks that tend to be difficult. This happens mainly because of the high number of numeric parameters, and specific technical details. Another issue that prevents the approaches from being replicated is the fact that the code developed is rarely made publicly available. In this essay we discuss and provide some guidelines to address these problems. Our goal is not to provide a unique, right answer, for these issues. Rather, we aim to promote a healthy discussion that can lead to new and innovative ideas and practices.

## 1 Introduction

When developing Machine Learning (ML) models for performing specific tasks (e.g., an Artificial Neural Network (ANN) for classification) the practitioner often undergoes a long and weary process of trial-and-error, where the structure and/or parameters of the model are continuously tuned in the search for the best performance. To avoid this, practitioners can resort to Evolutionary Machine Learning (EML), which uses Evolutionary Computation (EC) to optimise the ML models. Therefore, a population of individuals (each encoding a solution for the ML model) is continuously evolved, guided by a quality function that defines how well each model performs on solving the task.

However, according to Eiben and Jelasity [2] "verifying results found in the literature is in practice almost impossible". This happens for a number of reasons: the proposed approaches have specific and complex implementation details, parameters are not always clearly explained and detailed, and the code developed is not made publicly available. In addition, the evaluation of the methodologies is not performed in the same way, which makes the comparison between approaches even harder.

The critique by Eiben and Jelasity focuses on EC. But, similarly to what happens in EC, in ML models also have lots of parameters that need to be defined. Thus, the combination of both fields results in a bigger problem, since the number of parameters that must be clearly defined increases, making the reproduction of the results more challenging.

The main goal of this essay is not to propose one unique and right way of specifying the experimental setup. Instead, we want to promote discussion so that better practices can emerge and be used by authors in the area. In the upcoming sections we start by tackling reproducibility, and the comparison of approaches, respectively on Sections 2 and 3. Then, in Section 4, we focus on statistical analysis. To end, in Section 5, we discuss open questions and future directions.

## 2 Reproducibility

Reproducing the results obtained by EML methods is a difficult task. Since we are are combining two fields, namely EC and ML, there will be more parameters and technicalities than if considering only one in isolation. Therefore, if all the parameters that are used in the algorithms are not clearly pointed out, the reproduction of results and comparison of approaches is very hard.

One of the most common problems in EML is trying to understand how the benchmark dataset was partitioned. We advocate the use of three partitions: (i) train; (ii) validation; and (iii) test. The first two are used during the evolutionary process, and the last (test) must be kept out of it, and is used for measuring how well the evolved models perform beyond the data used for generating them, i.e., their generalisation ability. During the evolutionary process, the training set is used if the evolved ML model has a training phase (usual in supervised learning) for optimising numeric values (e.g., weights of ANNs), and the validation set is used for measuring the fitness value of the model, after training. However, if there is no need for a training phase, the train and validation sets can be merged, and consequently we only need two dataset partitions: validation and test.

We understand that different authors may use different names to mention the same sets that we defined above. The important message is that, the explanation of the dataset partitions, and what they are used for needs to be clear. More importantly, a percentage of the data must be kept outside the evolutionary process; otherwise the results are biased. Moreover, the way the partitions were created should be clear and reproducible.

Still focusing on the benchmark issues, whenever data augmentation, dimensionality reduction, or sampling techniques are used, both the method and its parameterisation must be provided. This applies for all other EC and ML parameters. Therefore we recommend that the experimental setup is reported on a table, in a clear and summarised way, avoiding the need for the reader to scan the entire article to search them. We divide the table into three distinct sections:

**EC** – specifies all the parameters directly concerned with the used evolutionary engine, such as, number of runs, number of generations, population size, crossover and mutation rates and parent selection mechanism;

**ML** – details the parameters of the evolved models, and the allowed ranges. Other information, such as the metric used for assessing the quality of the models, or the use of cross-validation can be also included here;

**Benchmark** – contains all the information regarding the benchmark partition, and when applicable, dataset sampling, augmentation and other parameters regarding any form of pre-processing.

More table sections may be required depending on the problem that is being solved. The same applies to the parameters contained in each section.

Despite a clear specification of the used parameters, the replication of results may still be difficult due to implementation and algorithmic details. Eiben and Jelasity [2] suggest that to avoid the previous a framework could be implemented, and used by all researchers, where only new features would need to be implemented. Although this standardisation of code is likely the ideal approach, it may not be feasible. Researchers use different programming languages, and there are many approaches to encode the same solutions. So, unless all variants are implemented and made available, researchers will tend to avoid standardisation. We defend that the easiest and most simple form of reproducibility would be to open up and share the code developed, by uploading it to repositories and include it within the paper.

## 3 Comparing Approaches

After specifying the experimental setup and conducting experiments there is the need to compare the approach with similar ones to acknowledge how it performs in the broader scope.

The first decision that needs to be made is concerned with the number of evolutionary runs that will be executed. Evolutionary Algorithmss (EAs) are stochastic search heuristics, and thus different runs can lead to very different solutions. Further discussion on the definition of the number of evolutionary runs is carried out in the next section.

ML results are typically presented in the form of tables, which report various performance metrics on the used models. In addition, in specific domains, e.g. ANNs, specific criteria concerning the structure of the models also tend to be presented: number of neurons, layers and connections. Whatever the chosen metrics are, what is important is to clearly define them, because those are the properties that are going to be used to compare one methodology with others. Further, we defend that a plot depicting the fitness evolution across generations should be presented, because a table does not allow one to check evolution and convergence speed, which may be important in problems where time is crucial.

Some authors just report the results attained by the best model found throughout evolution. However, as above stated, EML approaches have a stochastic behaviour. Therefore, showing just the best result does not capture the overall picture of the tested methodologies, and in some circumstances the best solutions can be deceiving outliers. By presenting the average of the best individuals found in each evolutionary run, along with its standard deviation it becomes possible to verify if the methodology consistently finds suitable solutions or not. The best result can also be presented separately, but never at the cost of discarding average ones.

Even though results can be shown in terms of averages, they provide little information. It is possible to have an approach A that, on average, is slightly superior to an approach B, but nonetheless the difference is insignificant, and thus the approaches can be considered equal in terms of performance. To effectively acknowledge the superiority of one method over another, based on empirical data, we must use statistical tests.

## 4 Statistical Analysis

When using EAs it is unlikely that two consecutive runs lead to the same results. Randomness is an important part of the evolutionary process, and thus the stochastic essence of the methods requires multiple repetitions of the experiments in order to gather enough experimental data to apply a sound statistical analysis. Next, we discuss how we think that the experimental analysis should be conducted, considering aspects like the number of runs, initialisation of the populations, and the statistical tests that should be used.

Before starting any experimental study one should define the number of runs, $N$, the hypothesis, $H$, that is to be tested, and the significance level, $\alpha$. The number of runs identifies the amount of executions of the algorithm. Different runs can generate distinct results; if for different runs the algorithm consistently gives similar results it is possible to state that it is robust. Thus, the larger the value of $N$ the better we can assess the robustness of the approach. Additionally, it also defines the size of the sample that will be used by the statistical methods. Therefore, we need a large value of $N$, which as a rule of thumb should not be lower than 30. Next, the hypothesis $H$, which is a statement that we want to assert as true, must be defined. The last step consists on the definition of the significance level $\alpha$. The significance level is used in the statistical tests as the cutoff value to reject the null hypothesis. Commonly used values for $\alpha$ are either the 0.05 or the 0.01. The lower the significance level, the more the data must diverge from the null hypothesis to be significant.

Once we have executed the methods, and gathered all the samples, we need to ascertain what distribution our data follows in order to decide which type of statistical test to use. Some are based on the assumption that the data follows a certain distribution. When this happens it is possible to use a set of statistical procedures called parametric tests. The most common assumption is that the distribution of the samples follows a normal distribution. To check if the samples follow a normal distribution it is possible to use two tests: Kolmogorov-Smirnov and Shapiro-Wilk. If the test is non-significant, it tells us that the distribution of the samples is not different from a normal, and thus the we can assume that probably the data is normal. After this, and before selecting any parametric test, we need to check if the variance is homogeneous.

The last part of the statistical analysis is concerned with the hypothesis testing and the reporting of the results. To test the hypothesis we can use two types of statistical procedures: (i) parametric, which assumes that the sample data comes from a population that follows a probability distribution based on a fixed set of parameters, e.g., a normal distribution; (ii) non-parametric, that makes no assumptions about the data distribution. There are several tests available in each category. We need to check which is the one that suits our assumptions the best. For example, if we are to compare two approaches A and B, with different initial conditions and with no assumptions about the distribution of the samples, we have to select a non-parametric test (most common situation in EML). Based on these assumptions the test that is most appropriated is the Mann-Whitney U test. For other scenarios and statistical procedures, please refer to [3].

Care must be taken with the interpretation of the word "significant", because even if the probability of the effect in our results occurs by chance is small (less than $\alpha$), it does not imply that the effect is of great importance. Insignificant and unimportant effects can be significant due to the large number of experiments conducted. So the question now is knowing how important an effect is. The solution to this problem is to measure and report the size of the effect that we are testing, known as effect size.

Table 1: Graphical overview of the statistical results with effect sizes.

| | | Dataset-A | Dataset-B | Dataset-C | Dataset-D |
|---|---|---|---|---|---|
| RMSE | Test | ++ | $\sim$ | +++ | ++ |
| | Validation | $\sim$ | $\sim$ | +++ | $\sim$ |
| Accuracy | Test | ++ | $\sim$ | +++ | ++ |
| | Validation | ++ | $\sim$ | ++ | $\sim$ |
| AUROC | Test | ++ | $\sim$ | +++ | ++ |
| | Validation | ++ | $\sim$ | +++ | $\sim$ |
| F-measure | Test | +++ | $\sim$ | +++ | $\sim$ |
| | Validation | ++ | $\sim$ | +++ | $\sim$ |

The effect size is a simple and standardised measure of the magnitude of the observed effect. Since it is a standardised measure it means that we can compare effects sizes across different studies that have different metrics. In the literature there are many methods to compute the effect size; the Pearson's $r$ correlation coefficient is one of the most widely used ones. One of the advantages is that it is constrained between 0 (no effect) and 1 (a perfect effect). Cohen [1] has made the following suggestions for a scale of the effects: small ($0.1 \leq r < 0.3$), medium ($0.3 \leq r < 0.5$), and large ($r \geq 0.5$). We recommend reporting the effect size in the form of a table, where the approaches are compared according to the following graphical overview: $\sim$ indicates no statistical difference between the compared approaches, and $+$ signals that approach A is statistically better than approach B. The effect size is denoted by the number of $+$ signals, where $+$, $++$ and $+++$ correspond respectively to small, medium and large effect sizes. A $-$ signals scenarios where approach A is worse than approach B. Table 1 shows an example following these guidelines.

## 5 Road Ahead

In this paper we have discussed multiple issues in experimental design, and how they affect EML. In particular:

- The reproduction of experimental results is nearly impossible. As EML merges the EC and ML fields, the number of parameters that needs to be set up is extremely high. The same happens to implementation details, which if not clearly specified, and if the code is not shared, make it almost impossible for other authors to replicate the obtained results. In that sense we propose the creation of a platform for sharing the benchmarks and obtained results, where the code may be made available;

- The reporting of results cannot be based only on the presentation of the best models. At least average values should be provided, along with the standard deviation, so that it is possible to analyse the consistency of the evolved models over the different runs;

- The comparison of different approaches using statistical procedures is essential. Most of the current published works do not rely on any statistical inference tools, or, when they do, the report of the analysis is not adequate, and very difficult to follow. In this work, we propose a recipe to fill this void by defining a set of guidelines that aim at improving and easing the comparisons between different works.

The current essay is by no means an extensive review of the literature. There are much more open questions that need to be answered. One of the most prominent ones concerns the increasingly necessity to find objective measures that are able to define what makes a good model. It is true that there are many performance metrics (e.g., accuracy or RMSE); however, when the results of two different methods are very close, should one choose a model that is more complex, despite the small increase in performance?

## Acknowledgements

This work is partially funded by: Fundação para a Ciência e Tecnologia (FCT), Portugal, under the grant SFRH/BD/114865/2016.

## References

[1] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences.* Routledge, 1988. ISBN 0805802835.

[2] Agoston E Eiben and Márk Jelasity. A critical note on experimental research methodology in ec. In *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, volume 1, pages 582–587. IEEE, 2002.

[3] Andy Field. *Discovering statistics using SPSS.* Sage publications, 2009.

# Data mining and social network analysis for personalized paper retriever: A Case study in ArXiv

André Luiz Pilastri[1]
andre.pilastri@fe.up.pt
Daniel F. Nogueira[2]
daniel.nogueira@fe.up.pt
Danilo Samuel Jodas[1]
danilojodas@gmail.com

[1] Faculty of Engineering - University of Porto, Portugal
Informatics Engineering Department
[2] Instituto de Ciência e Inovação em Engenharia Mecânica
e Engenharia Industrial,
Faculty of Engineering - University of Porto, Portugal

## Abstract

In this paper, the study and application of data analysis techniques for extracting information is proposed. The contribution of this work targets the process of identification of relevant literature from a collection of crawled documents. Novel functions, called social network features, are described and evaluated on documents crawled on ArXiv[1], to examine their relevance. The results highlight the data analysis process and the performance of the classification of the data mining algorithms used.

## 1 Introduction

The amount of data available in digital format on the World Wide Web (Web) is growing steadily. Due to scientific advances, scientists around the world continue to produce high amounts of research articles, which provide the technological basis for worldwide dissemination of scientific discoveries. There are many online digital libraries such ArXiv, ACM Digital Library[2], Google Scholar[3], Microsoft Research[4], and others that store research articles or their metadata. They have become a medium for answering questions such as: how research ideas emerge, evolve, or disappear as a topic; what is good measure of quality of published works; what are the most promising areas of research; how authors connect and influence each other; who are the experts in a field and what works are similar [1].

In particular, ArXiv is the largest source of high quality streamed science data. It is a repository with a collection of over one million open-access documents, including scientific papers in the fields of mathematics, physics, astronomy, computer science, quantitative biology, statistics, and quantitative finance. This way, ArXiv can be used to provide information that may be applied in data mining techniques to analyze research papers.

In this paper, we formulate the research article identification issue as a problem of relevant research articles from a collection of crawled documents from Arxiv. Moreover,the ability of a network representation of different bibliographic descriptive parameters in ArXiv to discriminate the relevance of a set of crawled documents is evaluated.

## 2 Case Study - Personalized Paper Retriever

This section summarizes some achievements in the process of automatic recommendation of scientific articles, and consists of many sub-tasks. In general, text classification is a problem divided into nine steps: data collection, text processing, data division, feature extraction, feature selection, data representation, classifier training, applying a classification model, and performance evaluation [4, 5, 6, 9].

In text classification, the first step is collecting data. The sample data used in this work consists in text that belongs to a limited scientic domain [7], i.e., according to a defined query (compressive + sensing + AND + image + reconstruction + OR + image+processing). According to [8] after setting goals required, a data mining process includes the definition the process objectives and materials necessary: describe the problem in a clear way; identify the data relevant to the problem at hand; ensure that the relevant variables for the project are not interdependent. More than creating a simple model, we wanted to compare how different algorithms could impact on the results and the knowledge over the data.

Following the established criteria, a manual scoring of a total of 305 obtained articles was conducted. The binary scoring of the relevant versus non-relevant articles was performed as a ranking of priority 1 versus priority 2, respectively, of each article. A total 166 articles with priority 1 were selected for dataset 1. A new set of features, which were extracted from the analysis of the social network of authors Table 1, was included in the dataset 2. Network Analysis is the field that studies the structure of a network in order to retrieve important information about its elements and interactions between them. Deciding the most important or central elements is a relevant aspect in the analysis of a network. This can be achieved through the use of centrality measures. The network centrality measurements that were used in this work in applications of research document analysis are presented in Table1.

Table 1: The list of Input Fields

| Input Fields | Description | Dataset | |
|---|---|---|---|
| | | 1 | 2 |
| Title | Title of article | √ | √ |
| TNR | Total number of reads | √ | √ |
| TND | Total number of downloads | √ | √ |
| TNC | Total number of citations | √ | √ |
| NRC | Total number of refereed citations | √ | √ |
| Similarity words | Frequency or occurrence of each word | √ | √ |
| Degree | The total number of edges connected with node | | √ |
| Betweenness | Calculate the betweenness centrality for each vertex and edge. | | √ |
| Closeness | Measure centrality | | √ |
| Pagerank | Calculate the PageRank of each vertex | | √ |
| Transitivity local | Measures the probability that the adjacent vertices of a vertex are connected. | | √ |
| Priority | Priority of article | 166 | 166 |
| | Total of articles | 305 | 305 |

- **Degree:** the simplest connectivity measurement is the node degree [2, 3], which corresponds to the total number of edges connected with node $i$. This measurement is defined for directed networks as $k^{(in)} = \sum_i a_{ij}$ and $k^{(out)} = \sum_j a_{ij}$ for in and out degree, respectively. If one considers the undirected and unweighted version of the network, the degree $K_i = \sum_j a_{ji} = \sum_j a_{ij}$ can be understood as the number of distinct bi-grams that a given words appears.

- **Betweenness:** Many recent studies of networks of various types have focused on network distance between nodes. This distance is defined as the number of hops along links in the network that one needs to make to move from one given node to another. To define this measurement, consider all paths connecting any pair of nodes in the network are followed via shortest paths [2]. The betweenness of a node $u$ is defined as being proportional to the number of paths that passes through node $u$. More specifically,

$$B_u = \sum_{ij} \frac{\sigma(i,u,j)}{\sigma(i,j)} \tag{1}$$

  where $\sigma(i,u,j)$ is the number of shortest paths between $i$ and $j$ that passes through node $u$ and $\sigma(i,j)$.

- **Closeness**: unlike the betweenness centrality, which is based on the number of shortest paths, the closeness centrality [3] uses the length of shortest paths. If $d_{ij}$ is the shortest distance between nodes $i$ and $j$, the closeness centrality is calculated as $C_i = V^{-1}\sum_j d_{ij}$.

- **PageRank:** is widely known to be part of the Google's web search [3]. The value of PageRank of vertex v, PR(v), is given iteratively by the relation:

$$Pr = \alpha AD^{-1}Pr + \beta 1 \tag{2}$$

  where $\alpha$ and $\beta$ are positive constants (conventionally $\beta = 1$), 1 is a vector $(1,1,1,\ldots)^T$ and $D$ is a diagonal matrix represented as

$$D_(ij) = \begin{cases} \max k_i^{(out)}, 1 & \text{if } i \neq j \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

[1] http://arxiv.org
[2] http://dl.acm.org
[3] https://scholar.google.pt
[4] https://www.microsoft.com/en-us/research/

The o PageRank considers a weighted sum of neighbors importance reflecting the neighbors degree. In this way, the relevance associated to a node is proportionally transferred to its neighbors.

- **Transitivity local:** is measures the probability that the adjacent vertices of a vertex are connected. This is sometimes also called the clustering coefficient. The local transitivity of an undirected graph, this is calculated for each vertex given in the vids argument. The local transitivity of a vertex is the ratio of the triangles connected to the vertex and the triples centered on the vertex. For directed graph the direction of the edges is ignored.

## 3 Experiments and Discussion

This section describes the infrastructure used to perform the experiments and also illustrates and discusses the obtained results. More than creating a simple model, we wanted to compare how different algorithms could impact on the results and the knowledge over the data. Using RapidMiner[5], the following five well-known algorithms for the prediction task were chosen: Decision Tree, Numerical ID3, K-NN, Naive Bayes and Rule Induction. For all five algorithms we have followed the same methodology for improving the model and check performance. Firstly, a normalization technique was applied to the training dataset. The main problem of this approach was that the dataset of relevant/non-relevant samples was not balanced, so the training set could be biased by the distribution of the data. To address this issue, a 70-30 split mechanism was used. Moreover, a stratified sampling technique was used to sample the training set, and considered in the performance validation step, for result comparison and parameters refinement. The results of the test are indicated in Table 2.

Table 2: Performance of classifiers using Dataset 1 and 2

| Classifier | Recall% | Precision% | Accuracy% |
|---|---|---|---|
| | D1$\|\|D2$ | D1$\|\|D2$ | D1$\|\|D2$ |
| Rule Induction | 29.03\|\|19.35 | 56.25\|\|85.71 | 68.13\|\|71.43 |
| ID3 Numerical | 41.94\|\|51.61 | 76.47\|\|76.19 | 75.82\|\|78.02 |
| K-NN | 48.39\|\|74.19 | 78.95\|\|76.67 | 78.02\|\|83.52 |
| Naive Bayes | 80.65\|\|77.42 | 43.86\|\|77.42 | 58.24\|\|84.62 |
| Decision Tree | 80.65\|\|80.65 | 75.76\|\|78.12 | 84.62\|\|85.71 |

The results indicated in Table 2, show the hit rate in class 1 (True Positive). It is possible to notice that the Naive Bayes and Decision Tree classifier achieved 80.65% of Recall in the D1 and D2 dataset, meaning a high correct predictability of the true positive cases. With Figure 1 we observed the resulting model as the tree output. The method of classification by decision tree works as a flow chart in a tree, where each node (not leaf) indicates a test done on the value. The links between the nodes represent the possible values of the top node test, and the leaves indicate the class (category) to which the record belongs.

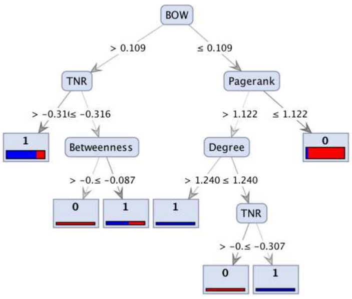


Figure 1: Result of decision tree of their D2 dataset

The main objective was to analyze the results of classification using features of co-authorship networks. The results indicated a significant improvement in Recall, whereas the top performing model was the Decision Tree classifier. This happens since some features of the co-authorship network, like Betweenness and Pagerank, were relevant to the classification process.

## 4 Conclusions and Future work

In this work, a model for recommending scientific articles applied to a dataset gathered from the ArXiv website was proposed. In general, building a large amount of labelled training data for text classification is a labour intensive and time-consuming task. This approach has promising results, mainly because it is suitable to be applied in all domains. In future work a comparison of this model with the one developed by [10] is pertinent. Furthermore, the results presented can be improved with further building and analysis a network of articles in order to find new features and look into the possibility of enhancing the proposed document retriever performance by using social networks.

## 5 Acknowledgments

The authors gratefully acknowledge "Conselho Nacional de Desenvolvimento Científico e Tecnologia - CNPq" for the scholarship provided during this research.

## References

[1] Cornelia Caragea, Jian Wu, Kyle Williams, Sujatha Das G, Madian Khabsa, Pradeep Teregowda, and C Lee Giles. Automatic identification of research articles from crawled documents, 2014.

[2] L. da F.and F. A. Rodrigues Costa, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007. doi: 10.1080/00018730601170527.

[3] Henrique F. de Arruda, Luciano da F. Costa, and Diego R. Amancio. Using complex networks for text classification: Discriminating informative and imaginative documents. *EPL (Europhysics Letters)*, 113(2):28007, 2016. doi: https://doi.org/10.1209/0295-5075/113/28007.

[4] Carlos A S J Gulo and Thiago R P M Rúbio. Text mining scientific articles using the r language. pages 70–81, Porto, 2015. 10th Doctoral Symposium in Informatics Engineering - DSIE 15.

[5] Carlos A S J Gulo, Thiago R P M Rúbio, Shazia Tabassum, and Simone G D Prado. Mining Scientific Articles Powered by Machine Learning Techniques. In *2015 Imperial College Computing Student Workshop (ICCSW 2015)*, volume 49, pages 21–28, Dagstuhl, Germany, 2015. doi: 10.4230/OASIcs.ICCSW.2015.21.

[6] Mohammad S. Khorsheed and Abdulmohsen O. Al-Thubaity. Comparative evaluation of text classification techniques using a large diverse arabic dataset. *Language Resources and Evaluation*, 47(2): 513–538, jun 2013. doi: 10.1007/s10579-013-9221-8.

[7] Chitu Okoli and Kira Schabram. A guide to conducting a systematic literature review of information systems research. *SSRN Electronic Journal*, 2010. doi: 10.2139/ssrn.1954824.

[8] David L Olson and Dursun Delen. *Advanced Data Mining Techniques*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[9] Timothy N. Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, jul 2012. ISSN 0885-6125. doi: 10.1007/s10994-011-5272-5.

[10] Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 448–456, New York, NY, USA, 2011. doi: 10.1145/2020408.2020480.

[5] https://rapidminer.com

# On the importance of Users's Features in Retweeting

Nelson Oliveira[1]
njoukov@student.dei.ic.pt
Joana Costa[12]
joana.costa@ipleiria.pt
Catarina Silva[12]
catarina@ipleiria.pt
Bernardete Ribeiro[1]
bribeiro@dei.uc.pt

[1] Center for Informatics and Systems
Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal

[2] School of Technology and Management
Polytechnic Institute of Leiria
Leiria, Portugal

## Abstract

Nowadays Twitter is one of the most used social networks with over 1.3 billion users. Twitter allows its users to write messages called tweets that can contain up to 140 characters. Retweeting is the key mechanism of information propagation. In this paper, we present a study on the importance of different user's features in predicting the probability of retweeting. The resulting retweet predictive model takes into account different types of tweets, e.g, tweets with hashtags and URLs, among the used classes. Preliminary results show there is a strong relation between specific features, e.g, user's popularity, and the retweet model.

## 1 Introduction

Twitter is a very popular micro-blogging social network founded in 2006. Twitter plays a role of news media when announcing breaking news [2]. In this social network a user can post a message (tweet), which can be shared (retweeted) by other users. Currently, Twitter has over 1.3 billion accounts, but it is estimated that only 550 million had ever made a tweet. Twitter is widely used by social media, brands and celebrities, which can influence their reputation.

Popularity prediction has been studied by multiple authors [5], [4], [6]. Predicting popularity can allow the users to better structure their tweets in order to understand how they can have higher possibilities of making a tweet that can achieve a higher popularity, which can be useful, for instance, to companies [3] that want to extend their popularity in order o increase their reputation and profits.

Some studies investigated which are the factors that are most important to popularity prediction. Suh *et al.* [1] and Janders *et al.* [4] showed that URLs and hashtags are have a positive influence, along with the number of followers (user's popularity), which is considered one of the most influencing features.

Other studies focused on building retweet prediction models. Petrovic *et al.* [5] built a retweet prediction model that tells if a tweet is going to be retweeted or not. Janders *et al.* [4] used Naïve Bayes and Logistic Regression to classify viral tweets and non viral tweets using different tresholds for virality, adding sentiment analysis related features. Hong *et al.* [6] used SVM with the Information Gain of the features to predict the popularity of a tweet using multiple classes.

## 2 Proposed Approach

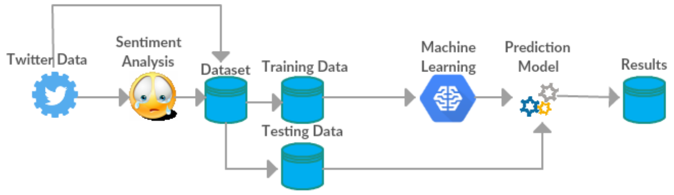


Figure 1: Retweet Prediction Model

The goal of this work is to build a retweet prediction model (Figure 1) and show how different quantity of text features (for example tweets with only 1 hashtag, tweets with less than 60 characters) affect its performance regarding different values of popularity. The text features that we tested were number of hashtags, number of URLs, and length of the text. We used 4 classes as Zhang *et al.* [6] to measure the popularity of a tweet. Class 1 contains tweets that have 0 retweets, class 2 contains tweets that have more than 0 and less than 10 retweets, class 3 contains tweets that have 10 or more retweets and less than 100, and class 4 contains tweets that have 100 or more retweets. Our model uses tweets that were collected with the Twitter Streaming API [1] and performs sentiment analysis of tweets, which were used as training data and testing data. The training data was generated using 2 sampling methods, and fitted to a Random Forests classifier. Finally, we tested the model with different testing sets.

## 3 Experimental Design

### 3.1 Dataset description

The dataset used in this work was gathered with the Twitter Streaming API, from 1st to 15th July 2016. The dataset has 12,470,144 (English) tweets. Which are distributed by the 4 classes in the following way: 8,684,496 tweets in class 1, 2,276,806 tweets in class 2, 999,511 tweets in class 3 and 509,331 tweets in class 4. We chronologically divided the dataset into two parts: tweets from the days 1 to 12 were used as training data, and the tweets from the days 12 to 15 as testing data.

### 3.2 Sampling Methods

For the sampling we used 2 sampling methods to select the training data, the first one we called **Class balancing** and the second one we called **Sub class sampling**.

#### Class balancing

Based on the unbalanced characteristics of our dataset the first sampling method that we have applied was to select randomly 300k tweets from each one of the classes, in order to have the same number of samples of each class which would not happen with random sampling.

#### Sub class balancing

In this approach we divided the classes into multiple sub classes, in order to avoid the unbalanced distribution inside each class, as tweets with higher number of retweets are scarser. This also happens inside of the classes for example if we look at the class 3 which represents the tweets with 10 or more retweets and less than 100, the first interval for which we count the number of tweets within that range (10-20) has almost 10 times more than the range(90-100), so when we randomly select 300k tweets from this class most of the tweets are inside the first interval, which can influence the prediction for the tweets that are near the edges of each class.

### 3.3 Features

The features that we have used in our study can be divided in two main groups, user features and tweet features. Most of these features can be extracted from the structure of the data directly when it is gathered from the Twitter Streaming API, but we have also added other extra features related to the tweet.

User features are the ones related to the author of the tweet. All of these features already have been used in other studies related to popularity

[1] https://dev.twitter.com/streaming/overview

Table 1: Results for F1-Score of the different experimental tests

| Class<br>Test | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Randomly selected tweets without sub class balancing** | 80.80 ± 0.10 | 39.45 ± 0.15 | 44.50 ± 0.21 | 58.60 ± 0.32 |
| **Randomly selected tweets** | 85.19 ± 0.07 | 45.74 ± 0.26 | 57.18 ± 0.31 | 78.18 ± 0.15 |
| **Tweets without hashtags** | 86.36± 0.05 | 43.22± 0.24 | 56.41± 0.21 | 78.93± 0.52 |
| **Tweets with 1 hashtag** | 78.65 ± 0.14 | 51.55 ± 0.11 | 58.21 ± 0.12 | 75.32 ± 0.26 |
| **Tweets with 1 or more hashtags** | 80.23 ± 0.24 | **52.12 ± 0.10** | 58.61 ± 0.16 | 75.72 ± 0.11 |
| **Tweets without URLs** | 83.58 ± 0.12 | 44.46 ± 0.23 | 56.38 ± 0.12 | 78.44 ± 0.36 |
| **Tweets with 1 or more URLs** | **87.72 ± 0.07** | 48.50 ± 0.12 | **59.38 ± 0.46** | 76.57 ± 0.36 |
| **Tweets with 0 to 60 characters** | 87.04 ± 0.08 | 39.77 ± 0.29 | 54.51 ± 0.34 | **80.34 ± 0.24** |
| **Tweets with 60 to 100 characters** | 85.36 ± 0.08 | 44.76 ± 0.16 | 56.40 ± 0.35 | 78.01 ± 0.25 |
| **Tweets with 100 to 140 characters** | 82.62 ± 0.09 | 50.83 ± 0.07 | 58.81 ± 0.12 | 77.08 ± 0.22 |

prediction. We used the following features: number of followers, statuses, favorites, number of times the user was listed, number of days of the account and if the user is verified.

Tweet features are the ones related to the tweet itself. We used the following features: number of hashtags, URLs, mentions, length of the tweets, number of words, is a tweet a reply, hour of the tweet timestamp. Besides these features we also added the number of videos, number of photos and GIFs. To our knowledge no previous studies used these features. We have also used features related to Sentiment Analysis, which were positive sentiment score (from 1 to 5) and negative sentiment score (from -1 to -5). These values were obtained using the framework SentiStrength [2] , which already was used by Janders *et al.* [4]. The last feature that we used here was a binary one, representing if a tweet contains a trend or not.

## 4 Results and Discussion

In Table 1, we present the experimental results. The first line contains the results when we tested the model with class balancing, and the remaining lines contain the results when we tested the model with sub class balancing, for the different test sets. In our experiments we used the 80:20 ratio for training and testing the model, so in this case since we used 300k of tweets for each class, respectively we tested the model with 300k tweets.

Comparing the global results when we used the sub class balancing, we noticed a big improvement in all of the classes especially in the last one, the reason why this happened was because, as we said the distribution inside each class is also unbalanced, meaning that divining each class into multiple sub classes helps in better selection of the training samples which lead in a performance of the model.

Analyzing the performance of the model regarding the different text features that we had tested, we showed that some features are more related to certain classes. As an example, shorter texts are more related to tweets that have no retweets (class 1) and tweets with more retweets (class 4), while longer tweets tend to be more related to the classes 2 and 3, where the performance of the model increases with the length of the tweets

In the case of hashtags we can say that tweets that do not have any hashtags are more easily identified when they will not get any retweets. These tweets of the class 0 can also be more related to the tweets that have at least one URL, meaning that regarding other features that we have used, when there is at least one URL the model performs better in respect of the class 0.

In the case when the tweet has one or multiple hashtags the classes 2 and 3 are the ones which have improvements in the performance, while the class 4 gets worse. This can seem strange since the presence of hashtags is associated with popularity. But we must also remember that the number of followers (user's popularity) plays one of the greatest roles when predicting retweet. So in this case we can think that popular users do not need hashtags in their tweets to have more retweets, which we can relate to the performance of the model when we did not use any hashtags regarding the class 4 where most of the tweets were made by users that are more popular than the ones contained in the other classes, which was better when testing with tweets that have at least one hashtag. In the classes 2 and 3 we can see that the model has the same behavior for these 2 classes, by that we mean when comparing the performance when testing with different values of the text features, to the performance of the model, the performance values for both of the classes either increases or decreases. This can mean that the number of retweets between the both ranges of these classes are affected in same way regarding to the different text features and their values.

## 5 Conclusions

In this work we have presented an analysis on the importance of different factors, as the user's popularity on a retweet prediction model, using 2 sampling techniques. The results of these techniques shows that dividing each class into multiple samples helps to increase the performance of the model. We also showed that the performance of the model is different regarding each class for tweets with different characteristics, which might suggest that there are some characteristics that are more related to a certain class, for example the best performance achieved by the model regarding the class 4 was when we used tweets with a short text (0 to 60 characters).

Possible extensions of this work can be such as, analyzing the performance of the model adding different user model profiles (i.e users with many followers, users with few followers and many favorites). This will help the different types of users to have their personalized probabilities of getting a certain "range" of retweets regarding the values of the features that they introduce in their text message.

## References

[1] P. Pirolli B. Suh, L. Hong and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the 2nd IEEE International Conference on Social Computing, SOCIALCOM*, pages 177–184, 2010.

[2] H. Park H. Kwak, C. Lee and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web, number 2 in WWW*, pages 591–600, 2010.

[3] Jean Burgess Merja Mahrt Katrin Weller, Axel Bruns and Cornelius Puschmann. 2014.

[4] Gjergji Kasneci Maximilian Jenders and Felix Naumann. Analyzing and predicting viral tweets. In *Association for Computing Machinery*, pages 657–664, 2013.

[5] S. Petrovic, M. Osborne, and V. Lavrenko. Rt to win! predicting message propagation in twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain*, 2011.

[6] Zhiheng Xu Yang Zhang and Qing Yang. Predicting popularity of messages in twitter using a feature-weighted model. In *International Journal of Advanced Intelligence*, 2012.

[2] `http://sentistrength.wlv.ac.uk/`

# Multi-model for wind speed forecasting

Nuno Conde
nftmconde@gmail.com
Departamentos das Engenharias - ECT
Universidade de Trás-os-Montes e Alto Douro

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

Paulo Salgado
psal@utad.pt
CITAB/ Departamentos das Engenharias - ECT
Universidade de Trás-os-Montes e Alto Douro

## Abstract

Wind power is weather dependent, variable and intermittent over various time-scales. For this reason, the estimation of value of electric energy produced from wind kinetic energy, even in short-term forecast horizon, is not an easy or precise task. So, designing an accurate forecasting model is an important contribution for a reliable large-scale wind power integration in the electric distribution systems. In this way, this work explores the use of a Multi-Model Kalman Filter (MMKF) to provide an estimate of the average hourly wind speed, in a 12 hours horizon. K-means algorithm was used to identify the structure of MMKF used to forecasting speed wind and tuned by several Kalman Filters (KF). Its accuracy is compared with the persistent model and with the standard statistical method AR. The experimental results have shown that MMKF is a suitable, robust and accurate model.

## 1 Introduction

As renewable energy, the utilization of wind power has been growing rapidly around the world, particularly in the last decade in Portugal. The strong investment in infrastructures was reflected in the fact that in the last year renewable energies accounted for 57% of the electricity supply in Portugal (REN - Rede Energética Nacional), ranging from hydroelectric power (28%) to wind (22%), biomass (5%) and photovoltaic (1.4%). Non-renewable production, however, accounted for 43% of consumption, divided between coal and natural gas[1].

The problem with wind power is that it is weather dependent, strongly variable and intermittent over various time-scales and, consequently, difficult to estimate the electric energy that will be produced [1]. The increased incidence of wind power in an energy network causes an increase of the unpredictability factor of energy production. This point can only be overcome if there is the ability to predict the power that will be injected into the electric distribution network [2]. This demand for prediction accuracy motivates researchers to propose accurate short-term forecasting models of wind power.

The main research in last decade is being carried out to obtain good wind speed forecasting systems based on physical methods [3] and statistical methods [4]. The first group has advantages in long-term prediction while the second group of methods do well in short-term prediction. Physical models consider the characteristics of geographic and local terrains and weather variables to estimate the future wind speed and generated power. The statistical methods use statistical tools to predict the wind speed and wind power. Recently, hybrid methods combining mathematical models for weather forecasting and statistical techniques have been proposed [5]. In this last group are included the autoregressive model (AR), the moving average model (MA), the autoregressive moving average model (ARMA), the autoregressive integrated moving average model (ARIMA) [6], the Box-Jenkins methodology and the use of the Kalman filter, KF, [7].

The main objective of this study is to test the performance of a forecasting multi-model to time series of wind speed. Its results are compared with persistent model (PM) and Autoregressive regressor model, AR. In previous work [8] the hybrid model with NN-Fuzzy Clustering has already been tested in the same way.

This work test a short-term forecasting multi-model in a parallel structure. It is the result of an identification process based on the analysis of historical time to find patterns in wind speed time series realized by K-means algorithm. Wind data collected on chosen meteorological stations are the main inputs for the classification process. The data used is the mean hourly wind speed values, in m.s$^{-1}$, measured on the location of Alvadia, district of Vila Real, Portugal. It covers a period from 1st of January 2007 to 17th of March 2009, representing 19368 h in total. This and other data is available from National System of Water Resources (SNIRH) [9]. The values of wind speeds from the previous hours are grouped into clusters according to their similarity.

In run time forecasting process, the clusters were used to select, in each time moment, the sub-model that will be used for forecasting the wind speed and its parameters are tuned using a KF method. The modelling steps of the proposed method are described as follows:

(1) Use K-means algorithm to decompose original time series into a number of different clusters. They are used to create sub-models of MMKF which can be used separately to predict and later on be recombined to get aggregate forecasting model.

(2) A sub-model is selected by cluster classifier for every sub-series. Use KF to tune the selected prediction model.

(3) Do aggregate calculation for prediction in sub-series to attain the final forecasting for original time series.

This paper is organized as follows. Section 2 presents a modified version of the K-means algorithm to be used in recognition of wind speed curves. This version is a fetch to adapt it to the time series data. In Section 3 we present the MMKF algorithm. Its performance is described in the next section with the presentation of the computational results, which are compared with others standard two methods, frequently used in forecasting wind speed. Finally, the last section presents the main conclusions of this study and about MMKF algorithm.

## 2 Clustering of time series

Let $V = [v_1, v_2, \ldots, v_k, \ldots, v_n]$ a series of data values of wind speed indexed in time order, here a sequence taken at successive equally spaced points in time. For each instant *k* a window of last 6 hours values of wind are grouped in a vector $x_k = [v_{k-1} \ldots v_{k-6}] - v_k \in R^6$. The *K*-means clustering is used to partition the *n* observations into *K* clusters, $C = \{C_1, C_2, \ldots, C_k\}$, in order that it minimize the within-cluster variance. This objective is formalized by a minimizing the following objective function:

$$\arg\min_{U} \sum_{i=1}^{K} \sum_{k=1}^{n} u_{ik} d_{ik}^2$$

where $d_{ik} = \|x_k - c_i\|$ is the distance value of observation $x_k \in C_i$ with the center of $i^{th}$ cluster, $c_i$. The algorithm assigning objects to the nearest cluster by distance, i.e., $u_{ik} = 1$ if $d_{ik} < d_{jk}, \forall j$ and $u_{ik} = 0$ for otherwise cases. Clustering is thus considered as a nonlinear optimization problem which is usually solved by an alternating scheme. The Fig. 1 shown the prototypes (centers) of *K*=18 clusters for time series of (differential) wind speed.

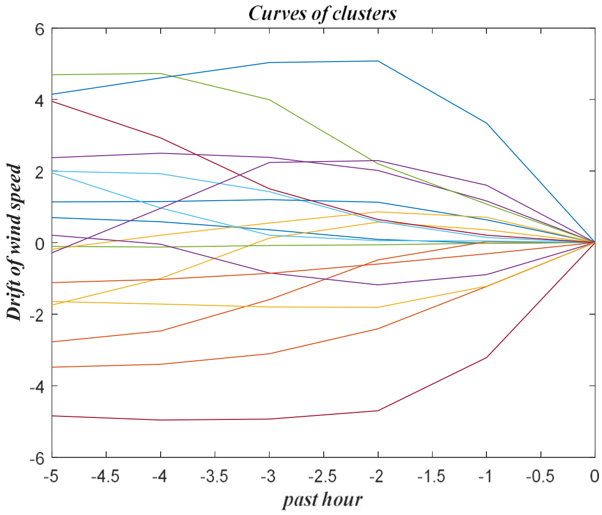


Figure 1: Prototypes (centers) of clusters

[1] https://www.ren.pt/pt-PT/media/comunicados/detalhe/ren_regista_novos_maximos_na_producao_de_energia_eolica/

## 3 The forecasting model

The MMKF is a multi-model structure learned by the *K*-means algorithm, with a number of sub-models equal to the number of clusters. Each cluster is associated to a sub-model.

The learning method is divided into three phases:

Step 1 - Structural Learning: initially, the data values of the wind from the last 6 hours are grouped into several clusters in an exclusive form based on the degree of similarity between the different data, grouping the values of wind according to their typical characteristics (clustering). The original time series is partitioned into a number of different clusters and expressed by the matrix *U* and the cluster centres, *c*, corresponding to a typical curve (curves prototypes) of wind speed.

Step 2 - Select sub-model and forecasting: For each wind profile $x_{k-1}$ select a sub-model according to the cluster classifier result. The result of this classification is expressed by the matrix *U*. Next, use MMKF to build the prediction models in the forecasting horizon.

Step 3 - Tuning MMKF parameters: Use KF for adjusting the weights of parameters selected model, grouped in vector *s*.

The KF is an algorithm that uses a model with parameter uncertainty and, a series of measurements observed over time, containing random noise and other inaccuracies, produce estimates of variables. It works in a two-step process: first step is for prediction where the KF produces estimates of the current state variables, along with their uncertainties. It involves the computation of the Kalman gain. This gain is then used in conjunction with the error in the prediction of the output, to correct the estimate of the state vector into its a posteriori form. The error covariance matrix is then computed for the updated or a posteriori state vector and a projection of it is obtained. An a priori estimate of the state vector is predicted for the next iterative cycle using the model of the system. The above steps are repeated iteratively. All equations used in this algorithm can be found in [10][11].

So, MMKF is a non-linear switched model selected by a classifier based on clustering. Each sub-model is a Linear Time-Varying System based on equations (1)-(2). Each parameter of the model is tuned by the KF. For each wind profile, $x_{k-1}$ , at time *k*-1, a sub-model is selected based on the cluster classifier. This is used to predict the wind speed $v_k$, at time *k*. The KF model assumes the true state at time *k*, $s_k$, is evolved from the state at (*k*−1) according to:

$$\boldsymbol{s}_{k|k-1} = A_k \boldsymbol{s}_{k-1|k-1} + w_k \ , \qquad (1)$$

where $A_k$ is the state transition model which is applied to the previous state $\boldsymbol{s}_{k-1|k-1}$ ; $w_k$ is the process noise which is assumed to be drawn from a zero-mean multivariate normal distribution with covariance *Q*. The matrix *A* is chosen to guarantee that (1) it is stable in terms of Lyapunov stability. In discrete domain a time *k* an observation (or measurement) $v_k$ of the true state $s_k$ is made according to

$$v_k = D_k \boldsymbol{s}_{k|k} + \upsilon_k \ . \qquad (2)$$

where *s* are the parameters of the forecasting model, $\boldsymbol{D}_k = \left[ v_{k-1} \ \mathrm{L} \ \ v_{k-6} \right]^T$ is the vector with the past 6th hours of observed wind speed values which maps the true state space into the observed space and $\upsilon_k$ is the observation noise which is assumed to be zero mean Gaussian white noise with covariance *R*. Equation (2) is the MMKF model.

Because of the algorithm's recursive nature, it can run in real time (here referred as online) using only the present input measurements and the previously calculated state, no additional past information is required, or with the last parameters of training process (offline model).

## 4 Results

Persistent and AR models are also used to forecast the wind speed. In the Persistent model, it is assumed that the forecast value of the time series is the last measured one, i.e., $v_t = v_{t-1}$ . The AR model structure is given by the equation $v_k + \sum_{i=1}^{p} a_i v_{k-i} = e_k$ which parameters ($a_i$, with $i$ =1, …, $p$) are estimated using variants of the least-squares method. The performance of the AR model was shown to be significantly better than the Persistent model. The proposed MMKF model was shown to be the best forecasting model. It was 64% better than the persistent model and 58% better than the AR model. In Figure 2 is plotted the response of all forecasting models for an interval period of approximately 400 hours.

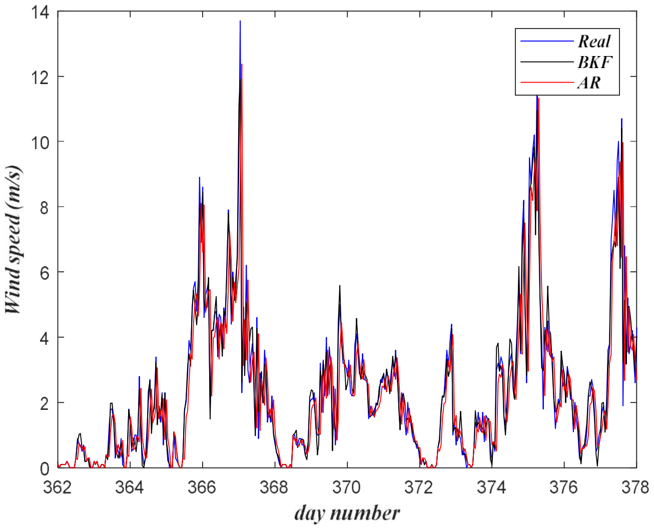


Figure 2: Wind speed forecasting of the MMKF (black line) and AR (red line) models; real wind speed wind (blue line) (test phase).

## 5 Conclusion

The comparison of various time series forecasting approaches on the mean hourly wind speed data indicated that MMKF based models outperformed the persistent and AR model. Its learning method proved to be equally viable and robust. Moreover, the experimental results make this model the most accurate with RMS errors in the range below of 0.2 ms$^{-1}$.

## 6 References

[1] M. Negnevitsky, and C. W. Potter, Very Short-Term Wind Forecasting for Tasmanian Power Generation, IEEE Transactions on Power Systems, vol. 21, no. 2, pp. 965 – 972, May 2006.

[2] M.C. Alexiadis, P.S. Dokopoulos, H. Sahsamanoglou and I.M. Manousaridis, Short-term forecasting of wind speed and related electrical power, Solar Energy 63(1), pp. 61–68, 1998

[3] Landberg L, Myllerup L, Rathmann O, Lundtang Petersen E, Hoffmann Jørgensen B, Badger J, Gylling Mortensen N. Wind resource estimation–an overview. Wind Energy 2003;6(3):261-271.

[4] Xiaochen Wang, Peng Guo and Xiaobin Huang, A Review of Wind Power Forecasting Models, Energy Procedia 12, pp. 770-778, 2011.

[5] Liye Xiao, Feng Qian and Wei Shao, Multi-step wind speed forecasting based on a hybrid forecasting architecture and an improved bat algorithm, Energy Conversion and Management, Vol. 143, 1 July 2017, Pages 410-430.

[6] E. Erdem, J. Shi, ARMA based approaches for forecasting the tuple of wind speed and direction, Applied Energy 88 (2011) 1405–1414

[7] Louka P, Galanis G, Siebert N, Kariniotakis G, Katsafados P, Kallos G, et al. Improvements in wind speed forecasts for wind power prediction purposes using Kalman filtering. Journal of Wind Engineering and Industrial Aerodynamics 2008;96(12):2348-2362.

[8] Paulo Salgado and Paulo Afonso, Hybrid fuzzy clustering neural networks to wind power generation forecasting, 14th IEEE International Symposium on Computational Intelligence and Informatics (CINTI 2013), 19–21 November, 2013, 359-363.

[9] National System of Water Resources (SNIRH) ; http://snirh.pt/.

[10] Paul Zarchan; Howard Musoff (2000). Fundamentals of Kalman Filtering: A Practical Approach. American Institute of Aeronautics and Astronautics, Incorporated. ISBN 978-1-56347-455-2.

[11] Hamilton, J. (1994), Time Series Analysis, Princeton University Press. Chapter 13, 'The Kalman Filter'.

# Prediction of client subscription to a term deposit

Hugo Fragata[1]
hugofragata@ua.pt
Marcos Pires[1]
marcosnetopires@ua.pt
Petya Mihaylova[2]
petya.tihomirova@gmail.com
Petia Georgieva[1]
petia@ua.pt

[1] University of Aveiro, DETI, Aveiro, Portugal
[2] Technical University of Sofia, Sofia, Bulgaria

## Abstract

Defining a successful marketing strategy to attract new clients that would subscribe bank term deposits is a challenging problem for any bank management. Several business intelligence techniques have been previously proposed to serve as decision support systems. Most of them are based on machine learning approaches to identify different clients profiles.
With this project we aim to explore the most frequently used methods and compare the performance of the classical (Support Vector Machine, Decision Trees, k Nearest Neighbor, Logistic Regression, Linear Regression, Random Forest) and advanced (Deep Learning) machine learning techniques. Our findings show that 1) Deep Learning (DL) is not the best option for this problem; 2) Classical methods exhibit similar quality and outperform DL; 3) Data preprocessing, such as normalization, replacement of missing data, data encoding and data dimensionality reduction (PCA, Factorial analysis) usually improve the model performance; 4) Factorial analysis reveals as a powerful technique to extract the most discriminative features of the telemarketing calls, namely call duration and client age. These findings match closely marketing management hypothesis.

## 1 Introduction

Prediction of the client profile that most likely would be interested by a bank product is a key issue within the banking industry. Several valuable tools to support client selection decision have been proposed [1], [2], [3]. In this paper we search for the most relevant approach applied to a dataset that is a mixture of bank client data and a direct telemarketing campaign through phone calls. Our objective is to analyze which of the studied machine learning (ML) classification algorithms is the most accurate in predicting if a client will subscribe or not to a term deposit from a selected set of attributes about the client, such as education and age and from the phone call characteristics, such as call duration.

## 2 Data Set

This study considers real data collected from a Portuguese retail bank, from 2008 to 2013. Data are composed of bank client information and the results of 41188 telemarketing phone contacts, [2]. The dataset consists of 16 attributes and is unbalanced, as only 11.3 % of the records are related with success.

**Categorical Attributes**: job (12 values), marital status (4 values), education (8 values), client's credit in default (binary), housing loan (binary), personal loan (binary), contact communication type ("cell" or "phone" values), last contact day of the week (1-7), last contact month of the year(1-12), poutcome (outcome of previous marketing campaign, 3 categorical values).

**Continuous Attributes**: client age, current account financial balance of the client, call duration, number of contacts performed during this campaign for this client, number of previous contacts, number of days that passed after the last time a client was contacted during a previous campaign.

## 3 Decision Support System

The decision support system consists of the following steps:

1) **Data preprocessing**

- Data Encoding: attributes with string values were converted to numeric values.
- Normalization (MinMax Scaler): scaling each feature to a given range.
- Noramlization (Standard Scaling): remove the mean and scaling to unit variance.
- Imputation: replacement of missing data with substitute values (for example the average of the column).

2) **Data dimensionality reduction**

- Incremental Principal Component Analysis (IPCA): builds low-rank approximation of the data, [4].
- Factor Analysis (FA): explorative method similar to PCA, [5].

3) **Classification**

The following classifiers were implemented:
- Classical ML algorithms: Linear Regression (LR) Logistic Regression (LogReg), K-Nearest Neighbors (kNN), Support Vector Machine (SVM) with Linear, Polynomial or Radial Basis Function (RBF) Kernel (RBF, Decision Tree (DT), Random Forest (RF).
- Advanced ML algorithms - Deep Neural Autoencoder (DNA), Fig. 1.

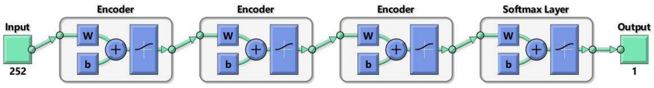


Figure 1: Deep Neural Autoencoder

## 4 Factor Analysis

Much like the cluster analysis of grouping similar cases, the Factor Analysis (FA) groups similar variables into the same factor, [5]. This process is usually referred as identifying latent variables. Due to its explorative nature, it does not distinguish between independent and dependent variables, it only uses the data correlation matrix.

In other words, FA reduces the information in a model by decreasing the dimension of the data set. This procedure can have multiple purposes. It can be used to simplify the data set, for example reducing the number of variables in predictive regression models. If factor analysis is used for these purposes, normally the factors are rotated after its extraction. FA considers several different rotation methods that ensure that the factors are orthogonal. Therefore, the correlation coefficient between two factors is exactly zero. For example, it totally eliminates problems of multicollinearity in regression analysis.

The most commonly used rotation method is *Varimax*. *Varimax* consists in an orthogonal rotation method (that produces independent factors) that minimizes the number of variables that have high loadings on each factor. The objective of this method is to simplify the interpretation of each one of the factors.Mathematically, the Factor Analysis is expressed as

$$\mathbf{X}_{(p\times1)} = \mathbf{L}_{(p\times m)} \times \mathbf{F}_{(m\times1)} + \varepsilon_{(p\times1)}$$

where each matrix represents the following:

$\mathbf{X}_{(p\times1)}$ - multivariate random vector with $p$ variables.

$\mathbf{F}' = [F_1\ F_2\ ...\ F_m]$ - **factors** vector with $m < p$ factors common to all $p$ random variables.

$\mathbf{L}_{(p\times m)}$ - matrix of **loadings** of the factors; the coefficient $l_{ij}$ represents the weight of the i-th variable to j-th factor.

$\varepsilon' = [\varepsilon_1 \ldots \varepsilon_p]$ - vector of errors.

The FA model considers that the variables could be grouped by their correlations. It is expected that when exists a high correlation between two variables they will be related to the same factor (in the sense that both will have high loadings on that factor).
Applying FA with standard Varimax rotation to the bank data revealed that the first two factors explain about 70% of the total data variability. Call duration and the client age have the highest loadings on Factor 1. Client education has the highest loading on Factor 2. Data visualization in Fig. 2 shows the high separability of the client's profiles based only on the first two features.

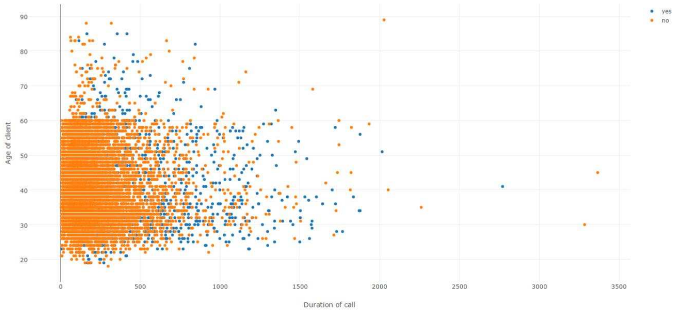


Figure 2: Data Visualization (call duration versus client age)

## 5 Experimental Results

In recent years, Deep Learning (DL) such as neural networks with multiple hidden layers, aka deep neural networks (DNN), has attracted great attention and won numerous contests in pattern recognition and machine learning, [6]. The DL underlying idea is to design a layer-wise algorithm, where each layer can learn features at a different level of abstraction. Hence, the user does not need to spend too much time in searching how to represent the data into the feature space. DNNs seem useful for solving classification problems with complex nonlinear data. One of the objectives of this work is to verify this hypothesis.
We applied Deep Neural Auto-encoder (DNA) with 3 hidden layers (with 50, 30, 10 nodes respectively), Fig. 1. The hidden layers are first pretrained in an unsupervised fashion using auto-encoders, [6]. The output layer is trained in a supervised fashion using labels for the training data. We have used softmax function as a processing unit in this layer. Final retraining of the whole multilayer network is performed on the training data in a supervised fashion applying standard backpropagation. This step is referred as fine DNA tuning.

Data were randomly divided into training set (60 %), cross validation set (20%) and testing set (20%). On Figs. 3, 4 and 5 are summarized the performance (in terms of accuracy on test data) of the best classical and the best DL classification models. In the experiments we combined the classifiers with different preprocessing steps, mentioned above. Both classical and DL classifiers may improve their performance after normalization and imputation. Our results show that the classical ML methods outperform DL models. SVM with RBF Kernel and imputed normalized data (Fig. 3) is the best performing method (90.1% accuracy), followed by the SVM with linear kernel and normalized data (Fig. 4).
The confusion matrix of the best performing model (Fig. 6) shows that the proposed system is able to advise with high confidence regarding clients that potentially will not subscribe a term deposit, however it is not that confident in recognizing the clients that may accept the bank proposal. This is clearly the effect of working with unbalanced dataset.

## 6 Conclusion

Building a client profile model before applying a marketing policy is valuable not only in banking industry but for any business. Our findings suggest that the feature selection engineering is a more promising approach than leaving it to a DL data representation. Particular emphasis needs to be paid on data normalization. In contrast to IPCA, where the principal components does not hold physical interpretation, FA relates the transformation factors and loading with the most discriminative features and thus provides an intuitive feature selection.

| Linear Support Vector Machine | 0.216445916115 |
|---|---|
| MinMax Scaled | 0.901434878587 |
| Standard Scaled | 0.900662251656 |
| Standard Scaled with IPCA | 0.886754966887 |
| Imputed LSVM | 0.844370860927 |
| Imputed MinMax Scaled | 0.899337748344 |
| Imputed Standard Scaled | 0.899337748344 |
| Imputed Standard Scaled with IPCA | 0.886754966887 |

Figure 3: Accuracy (on test data) of SVM with linear kernel

| 16RBF SVM | 0.886865342163 |
|---|---|
| MinMax Scaled | 0.899337748344 |
| Standard Scaled | 0.906622516556 |
| Standard Scaled with IPCA | 0.888631346578 |
| Imputed 16RBF SVM | 0.886865342163 |
| Imputed MinMax Scaled | 0.899337748344 |
| Imputed Standard Scaled | 0.906843267108 |
| Imputed Standard Scaled with IPCA | 0.887306843267 |

Figure 4: Accuracy (on test data) of SVM with RBF kernel

| Deep Neural | 0.339404 |
|---|---|
| MinMax Scaled | 0.886755 |
| Standard Scaled | 0.886424 |
| Standard Scaled with IPCA | 0.886755 |
| Imputed Deep Neural | 0.840177 |
| Imputed MinMax Scaled | 0.886755 |
| Imputed Standard Scaled | 0.886755 |
| Imputed Standard Scaled with IPCA | 0.886755 |

Figure 5: Accuracy (on test data) of DNA

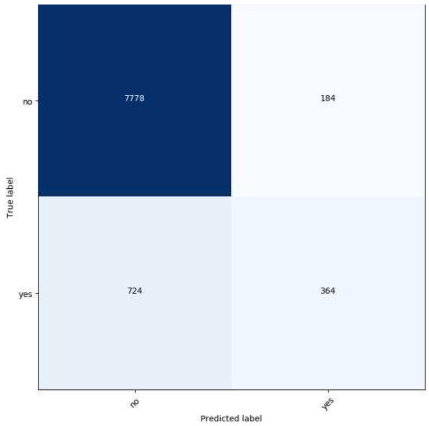


Figure 6: Confusion matrix of the best performing model (SVM with RBF Kernel and imputed normalized data )

## References

[1] http://media.salford-systems.com/video/ tutorial/2015/targeted _marketing.pdf

[2] S. Moro, P. Cortez, P Rita, A data-driven approach to predict the success of bank telemarketing Decision Support Systems, 62 (2014) 22-31.

[3] P. Kotler, K. L. Keller, Framework for Marketing Management, 5th edition Pearson, 2012.

[4] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157âĂŞ1182.

[5] L. Tucker and R. MacCallum. Exploratory Factor Analysis. University of Illinois and Ohio State University, 1997.

[6] G. Hinton, S. Osindero, and Y.-W. The. A Fast Learning Algorithm for Deep Belief Nets. Neural Computation, 18, 1527âĂŞ1554, 2006.

# Poster Session II

# Registration of Breast Surfaces Before and After Conservative Treatment of Cancer

Sílvia Bessa
silvia.n.bessa@inesctec.pt
Jaime S. Cardoso
jaime.cardoso@inesctec.pt
Hélder P. Oliveira
helder.f.oliveira@inesctec.pt

INESC TEC
Porto, Portugal

## Abstract

Surgery planing of breast cancer interventions is gaining importance among physicians, who recognise value in discussing the possible aesthetic outcomes of surgery with patients. Research is being propelled to create patient-specific breast models but breast image registration algorithms are still limited, particularly for the purpose of matching pre- and post-surgical data of patient's breast surfaces. Yet, this is a fundamental task to learn prediction models of breast healing process after surgery because pre- and post-surgery data has to be aligned. In this paper, a coarse-to-fine registration strategy is proposed to match breast surface data acquired before and after surgery. Methods are evaluated in their ability to register surfaces in an anatomically reliable way, and results suggest proper alignment adequate to be used as input to train deformable models.

## 1 Introduction

The choice of a registration algorithm generally depends on characteristics such as accuracy, computational cost, but most of all, it heavily depends on the data used. When aligning surfaces to learn breast deformation models from real data, non-rigid methods should be avoided. Pre and post-surgical data can actually differ in volume and shape, and these differences should remain intact after registration because models have to capture those transformations. Moreover, nipple should not be used to match surfaces, because sometimes it is lost during surgery.

Among rigid registration techniques, Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) have been used to compute an initial geometric alignment of data, but fine registration is commonly accomplished with the Iterative Closest Point (ICP) or its variants [1]. One of the few works that addresses the registration of breast surfaces is the work of Schuller *et. al* [6], which describes a multi-step procedure to find correspondences between point clouds (PCLs) of pendant and compressed breasts. PCLs were obtained from Computerized Tomography (CT) data, and a Finite Element Model was applied to simulate the compression of breasts. The initial correspondences between pendant and compressed data was found using ICP, but only applied to nipple and breast regions near the torso. Then, the symmetric nearest neighbour procedure proposed by Papademetris *et. al* [5] was used to identify point correspondences between the remaining volume of the breast, and propagate the initial registration. In this work (adapted from [2]), the problem of registering the 3D surface data of pre- and post-operated breasts is addressed, using rigid registration methods. Results suggest that proper alignment of surfaces is accomplished, enabling the use of real data to learn breast deformations models useful for surgical planing.

## 2 Proposed Methodology

To register pre- and post-surgical surfaces of breasts, a coarse-to-fine registration strategy is proposed. In the coarse registration stage, breast surfaces are geometrically aligned only by translating their centres of mass to the origin of the coordinate centre. A strategy based on fitting a plane to surface is also explored to determine the orientation between the two point clouds, and compared with PCA alignment. The fine registration of data is accomplished with ICP. Figure 1 shows an example of pre- and post-surgical data that has to be registered before learning healing process models, evidencing the differences in breast volumes and shapes. During surgery, the tumour and a margin of healthy tissue are removed, which reduces the volume of the breast. Then, healing process occurs which lasts about one year before the breast stabilises in its new shape. The tissues usually retract to adapt to the new internal structure, causing the breast to go up. This translates to deformations that have lower impact in upper profiles of the breast [3].Therefore, for modelling the breast shape changes after the healing process, some constraints have to be imposed on the alignment of pre- and post-surgical data. Instead of using all volume of the breast, one can argue that top profiles are more reliable to compute the transformation between the two PCLs, while preserving the expected anatomical behaviour. In fact, this was the followed strategy: both coarse and fine registration techniques were applied using different top profiles of the breast, and results were compared with the registrations using all volume of the breasts. Another challenge arises from limitations of the breast segmentation strategy [4]. Although registration is being computed using only the top areas of the breast, there is no guarantee that the segmentation algorithm outputs exactly the same surface from data acquired at different instants of time. Figure 2 shows an example of pre- and post-breast segmentations of a non-operated breast. Despite the natural shape of the breast did not change, a different segmentation resulted from data acquired before and one year after surgery. To overcome this limitation, and for each breast, different areas of the top profile are used to register pre- and post-data. In detail, percentages between 20% up to 100% were considered to define the top profile of the breasts (in the coronal view), with increments of 10%.

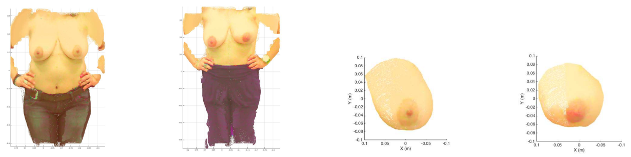

(a) Pre Torso (b) Post Torso (c) Pre (d) Post

Figure 1: PCLs of the torso and operated breast (patient's left) before and after surgery. Breast data was centred for visualisation purposes.

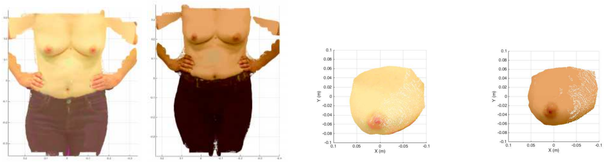

(a) Pre-Surgery Torso (b) Post-Surgery Torso (c) Pre-Surgery (d) Post-Surgery

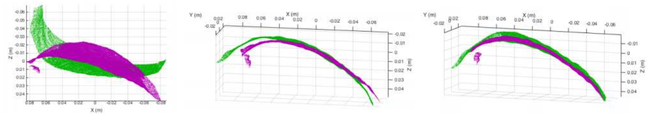
Figure 2: Differences in the segmentation of a non-operated breast PCL before and after surgery. Breast data was centred for visualisation purposes.

## 3 Experimental Results

The proposed methodology was evaluated on a real database containing breast PCLs of 30 patients, using euclidean and Hausdorff distances, calculated in both directions (Pre → Post, and Post → Pre). Data was acquired with one year interval, but no ground truth is available for assessing the quality of the registration procedure. As consequence, an exploratory analysis of the results is conducted instead, in which different percentages of the breasts are used to evaluate the registration strategy. In detail, after registration, the distances between the two PCLs are computed using equal percentages of pre- and post-breast data, between 20% and 100% of the breast points, with increments of 20%. Average and Standard Deviation of these distances are then used to search for the extents of pre-

and post-data that should be used to compute the transformation between the two PCLs. The idea is that the best registration strategy is the one that aligns data with small (low average) and similar errors (low standard deviation) in all portions of the breast.

In the proposed methodology, registration is constrained to top profiles of breasts. The top profiles of breasts are brought close in a coarse registration stage that considers three strategies: (i) simply centre PCLs based on the centroids of the top profile of the breast, or (ii) find the rotation angle between pre- and post-surgery data using either (iia) PCA axes or (iib) the fitting planes. Results show that despite being commonly used, PCA can produce unacceptable results when the main variance axes do not share the same direction in both PCLs (Fig. 3(a)), and PCA results were discarded from subsequent analysis. On the other hand, no significant differences ($p-value = 0.9364$, $IC = 5\%$) are found between the initial alignment based on centroids or fitted planes, when combined with the fine registration step using ICP. Figures 3(b) and 3(c) show the registration results of aligning pre- and post- top profile of operated breast of Fig. 1, with strategies (iia) and (iib), respectively. Note that registration results are better when using fitted planes, but the differences between centroid and plane alignment decrease when higher percentages of top breast profile are used. Yet, one can argue that a coarse fitting based on the fitting of planes to surfaces is more robust to data variability, and such strategy is advised instead of simply centring data.



(a) PCA (b) Centroids (c) Plane

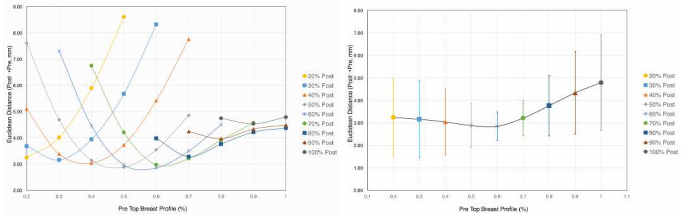
Figure 3: Coarse-registration results with pre- (green) and post-surgery (purple) top profiles ( 30%) of the breasts from Figure 1.

Different combinations of pre- and post-surgery top breast areas were also considered to tickle the problem of possible different breast segmentations. However, results showed that pre- and post-surgery top areas of breasts that differed more than 30% lead to poor results, and such combinations were no longer considered.

Regarding the overall performance of the proposed methodology, the values of distances computed in both directions were compared to infer if pre- and post-surgery breasts had different sizes. Although no significant differences were found between average euclidean distances ($p-value = 0.0889\%$, $IC = 5\%$), Hausdorff distances differed according to direction ($p-value = 0.0255\%$, $IC = 5\%$). An additional one-sided t-test further revealed that Hausdorff distances computed in $Post \rightarrow Pre$ direction have higher values than in the opposite direction ($p-value = 0.0128\%$, $IC = 5\%$), which confirms the expected: pre-surgical breasts are usually bigger than post-surgical ones.

Analysing the average euclidean distances for all combinations of pre- and post-surgery top breast areas, computed in $Post \rightarrow Pre$ direction, there is a trend where the combination of equal profiles of pre- and post-surgical breasts have smaller distances (Fig. 4(a)), in series using up to 60% post-surgery breast areas. For the remaining series, though, the optimal value is found when post-surgery top areas are combined with pre-surgery top areas with less 10%. Note that when a higher area of post-surgery breast is used, one increases the probability of calculating registration with surfaces of different sizes: breast healing process usually affects more the lower profiles of the breast. As consequence, increasing the amount of breast for registration increases the chances of getting larger distances between pre- and post-surgery breasts after transformation.

Figure 4(b) shows the results when equal portions of pre- and post-surgery top breasts are used. The smaller average euclidean distance is found when 60% of pre- and post-surgery breasts are considered in the registration, also having the low standard deviation value. The largest distances, both average and standard deviation values, are found when full breasts are used. In comparison, the average distance decreases when using low portions of breasts, but high values of standard deviation are found. This was the expected behaviour and what motivated the search for the best combination of pre- and post-surgery top breast areas in the first place. When using small portions of the breast, the top profiles of the breast can be very well aligned, but there is no restrain on how the remaining portions of the breasts are aligned, which causes high standard deviations. On the other hand, when full breasts are used, breasts are aligned by their lower profiles, and distances between the top areas increase.



(a) All combinations of pre/post percentages. (b) Equal pre/post percentages

Figure 4: Average euclidean distances ($Post \rightarrow Pre$ ) for different combinations of pre- and post-top profile percentages.

## 4 Conclusions and Future Work

The registration of pre- and post-surgery data of breasts is an essential task to learn breast healing process models useful for surgery planing. However, this task poses some challenges, namely the lack of landmarks before and after surgery, in addition to differences in size and segmentation. To overcome these limitations, a registration strategy that uses only top profiles of breasts is proposed, and different combinations of top breast areas were explored. Results suggest that such strategy outputs good pre- and post-surgery alignments of breasts, with the better results obtained when similar and intermediate portions of pre- and post-surgery top breast areas are used. To best of knowledge, this is the first work addressing the problem of registering data of breast before and after surgery.

## 5 Acknowledgements

This work is financed by the ERDF - European Regional Development Fund through the Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement and through the Portuguese National Innovation Agency (ANI) as a part of project BCCT.Plan - NORTE-01-0247-FEDER-017688, and also by Fundação para a Ciência e a Tecnologia (FCT) within PhD grant number SFRH/BD/ 115616/2016.

## References

[1] Ben Bellekens, Vincent Spruyt, Rafael Berkvens, and Maarten Weyn. A survey of rigid 3d point cloud registration algorithms. In *AMBIENT 2014: the Fourth International Conference on Ambient Computing, Applications, Services and Technologies, August 24-28, 2014, Rome, Italy*, pages 8–13, 2014.

[2] Sílvia Bessa and Hélder P Oliveira. Registration of breast surface data before and after surgical intervention. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 226–234. Springer, 2017.

[3] Maria João Cardoso, Helder Oliveira, and Jaime Cardoso. Assessing cosmetic results after breast conserving surgery. *Journal of Surgical Oncology*, 110(1):37–44, 2014.

[4] Hélder P Oliveira, Jaime S Cardoso, André T Magalhães, and Maria J Cardoso. A 3D low-cost solution for the aesthetic evaluation of breast cancer conservative treatment. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 2(2):90–106, 2014.

[5] Xenophon Papademetris, Albert J Sinusas, Donald P Dione, R Todd Constable, and James S Duncan. Estimation of 3-d left ventricular deformation from medical images using biomechanical models. *IEEE Transactions on Medical Imaging*, 21(7):786–800, 2002.

[6] Daniel R Schuler III, Jao J Ou, Stephanie L Barnes, and Michael I Miga. Automatic surface correspondence methods for a deformed breast. In *Medical Imaging*, pages 614125–614125. International Society for Optics and Photonics, 2006.

# Enhanced Cerebral Vascular Segmentation with Harmonic Constraints

Jorge Miguel Silva[1]
jorge.miguel.ferreira.silva@ua.pt
Augusto Silva[1]
augusto.silva@ua.pt
Pedro Vilela[2]
ferrovilela@hospitaldaluz.pt

[1] IEETA - Institute of Electronics and Informatics
Engineering of Aveiro
Aveiro University, Campus do Santiago
[2] Hospital da Luz

## Abstract

Stroke is the main cause of complex disability and the second leading cause of death worldwide. A fast diagnosis is crucial for avoiding patient death or after effects. In the emergency service, the cerebral Computed tomographic angiography is one of the first diagnostic lines often accompanied by perfusion studies. As such, in this paper we propose a novel cerebral vascular segmentation methodology based on the perfusion studies with voxel based harmonic constraints.

## 1 Introduction

Stroke is the largest cause of complex disability in adults and the second leading cause of death worldwide [2]. Every year, there are 15 million stroke incidents of which only two thirds manage to survive [1, 4]. It is described by brain cell death resulting from blood flow blockage or rupture of a vessel in the brain. A rapid and correct diagnosis is critical to avoid patient death or after effects.

One of the first diagnostic procedures used to evaluate if a patient had a stroke is to perform a cerebral Computed tomographic (CT) angiography often comprising perfusion studies. A CT perfusion study requires fast imaging technologies and provides a series of contrast enhanced brain volumes along a series of predetermined time points. After iodinated contrast medium has been administered intravenously to the patient [7], it is possible to assess visually and computationally the various flows that travel through the cerebral vascular structure.

With this in mind, in this paper we decided to investigate the possibility of computationally performing a segmentation of the vascular structure considering the intrinsic dynamics of the perfusion studies. Each perfusion study is composed by 24 volumetric datasets (512x512x81x24 voxels), distributed typically along 46 seconds.

## 2 Methods

The segmentation pipeline can be divided into 3 major steps:

- 3D Registration of the volumes to compensate for the patient's involuntary head movements along the time course of the perfusion studies.
- The raw Segmentation of the study to obtain the mask of cerebral vascular structure and apply it to obtain the flow of contrast along the blood vessels.
- Finally, perform a harmonic decomposition on the masked CT for each voxel of the cerebral vascular structure to refine the segmentation process based on the relative harmonic weighting of the Fourier components.

In order to implement this pipeline, the Mevislab framework[1] was firstly tested. However, due to difficulties in automating the solution, we chose to use Python 3 and some of its libraries to implement the registration and segmentation steps leaving the visualization burden to Mevislab.

### 2.1 3D Registration

Co-registration is the process of trying to spatially align one dataset with another. In order to align a given volume $V(x)$ to a fixed volume $V_F(x)$, a spatial transformation $T(x)$ is required. This transformation can be: rigid, hybrid or completely non-rigid. As shown in Equation 1, the goal of the transformation $T(x)$ is to minimize a Cost function $C(x)$, which is determined by a similarity metric between a given volume $V(x)$ and the fixed volume $V_F(x)$. The higher the similarity between volumes, the lower the cost function $C(x)$ value assuming a monotonic behavior.

$$\hat{T}(x) = arg\min_{T} C(V_F(x), V(T(x))) \tag{1}$$

Several methodologies were tested to perform co-registration, however we chose an affine transformation to register the 3D volumes since it better minimized the cost function $C(x)$. The similarity metric was the Mutual Information [6] and the optimization strategy implemented was similar to that used by ANTS [3]. To implement the solution we used the dipy[2] Python library.

Firstly we created a similarity metric by specifying the number of bins to be used in the discretization of the joint and marginal probability distribution functions (PDF) - the selected value was 32 - and the percentage of voxels to be used for computing the PDFs. All voxels were used in order to achieve the best registration possible. Furthermore, in order to avoid getting stuck at local optima, and thus to accelerate convergence, we opted to use a multi-resolution strategy by building a Gaussian Pyramid with the default settings. These settings consist of 3 resolutions with 10000 iterations at the coarsest resolution, 1000 iterations at the medium resolution and 100 iterations at the finest. Additionally, in order to build the Gaussian Pyramid, the fixed volume was firstly smoothed at each level of the pyramid using a Gaussian kernel. Next, we defined the sub-sampling factors so that the coarsest image had a quarter of the original size, the middle resolution had half of the original size, and the image at the finest scale had the same size as the original image. Finally, we used these configurations to instantiate the registration class, and perform the Affine Transform (translation, rotation, scale and shear).

### 2.2 Segmentation

After the 3D registration, segmentation of the blood vessels took place. For this process we selected the intra-cranial region and performed the segmentation of the vascular structure. Figure 1 shows the different stages of this segmentation [5].

In order to select the intra-cranial space, we looked at the Hounsfield Scale, the standard reference scale used for measuring radiodensity in CT scanning. As shown by Table 1, the range of bone tissue goes from 200 to 3000 HU. On the other hand, blood with contrast ranges from 100 HU to 500 HU.

| Tissue | Hounsfield Units (HU) |
|---|---|
| Bone | 200 - 3000 |
| Blood with Contrast | 100 - 500 |
| Blood | 40 |

Table 1: Table of Hounsfield Units for different tissues

Since our first step was to select the intra-cranial region, we started by applying a threshold to all the volumes of the exam. A threshold at 500 Hounsfield Units (HU) was applied in order to detect bone and ignore other soft tissues. To the resulting binary image, a 3D morphological opening operation with a spherical structural element was performed to eliminate small objects in the intra-cranial region of the volume. Next, the volume was labeled and a mask constructed by selecting the labeled intra-cranial region. This mask was applied to the original registered examination, thus creating an exam containing only the interior of the skull.

[1] http://www.mevislab.de/

[2] http://nipy.org/dipy/

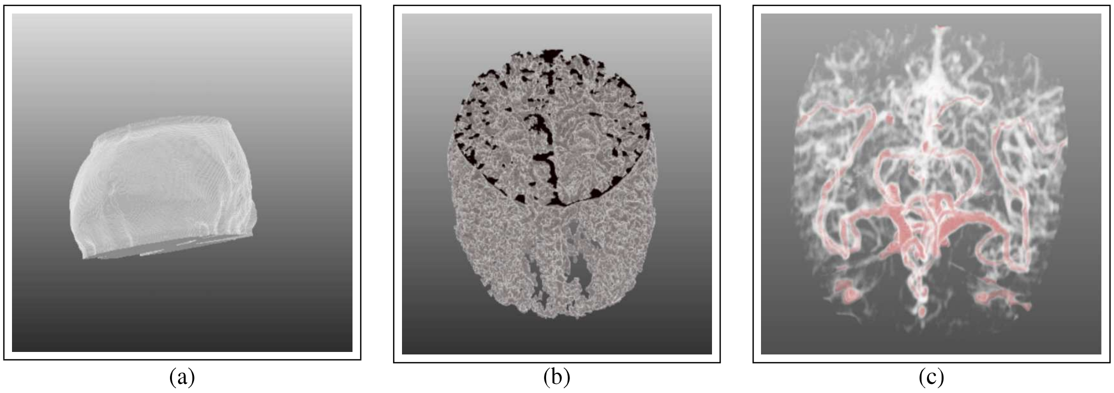


Figure 1: Different steps of Segmentation: (a) Intra-cranial Mask; (b) Vessel structure Mask; (c) Segmented Vessels for a specific time point.

After having the intra-cranial region selected we moved into vessel segmentation. We started by applying a threshold of 100 HU to the selected intra-cranial volumes of the exam, since contrast intensity in blood vessels ranges from 100 to 500 HU. The result was a masked exam where each of the volumes has the vessels containing contrast segmented. To obtain a complete mask of the vessel structure where contrast flows, we performed a logical disjunction operation for all the volumes to obtain a 3D array of all blood vessels where the contrast has passed. This mask was then applied to all time points of the exam and thus we obtained the vessel structure of the patient's brain with contrast flow for each time point.

## 2.3 Spectral Analysis

When analyzing pixel intensity for the vessel structures, we obtained an assumed periodic signal, as seen in Figure 2. As such, by selecting the pixels that possessed this behavior and removing the others, we could improve the image prior to segmentation.

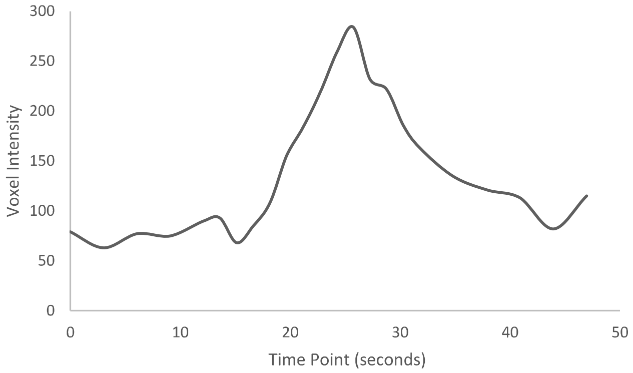


Figure 2: Example of pixel intensity over the duration of the exam for a pixel belonging to the vascular structure.

For that purpose, for each of the pixels in the time space a fast Fourier transform (FFT) was determined. The FFT algorithm computes the complex discrete Fourier transform (DFT), hence it results in a sequence of complex numbers of form $X_{re} + jX_{im}$ from where the Magnitude and Phase were determined.

By analyzing the results, we determined that the first harmonic was the most significant for representing contrast flow over the vessels. This first harmonic is presented in Figure 3, and it serves as the refined mask to filter noise present from the previous segmentation.

## 3 Discussion and Conclusion

In this paper we present a novel solution for the segmentation of blood vessels in the brain when performing a Cerebral CT Angiography comprising perfusion studies. We implemented a three step pipeline, consisting of 3D registration, raw segmentation and spectral analysis for a more accurate segmentation. The 3D registration allowed us to compensate for the patient's head movement. This was essential since without it, the vessels created by the mask during the segmentation process and the spectral analysis would not be accurate. The raw segmentation stage was performed considering the preference for over-segmentation rather than under-segmentation, to avoid losing pixels belonging to the vascular structure. The final stage of segmentation was carried out using magnitude's first harmonic. As such, our methodology follows a common segmentation strategy which is afterwards enhanced by the information of the flow dynamics associated to the first harmonic of the voxel based Fourier decomposition. The methods were applied for a total of 5 case studies. Although a visual analysis shows that a good segmentation was possible, there is still the need to test this methodology in a larger dataset.

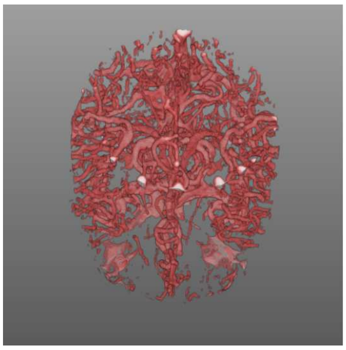


Figure 3: Mask obtained by the Magnitude's first harmonic.

## Acknowledgments

Jorge Miguel Silva is funded by the Research grant Project CMUP-ERI/ICT/0028/2014 - SCREEN-DR.

## References

[1] Stroke Statistics | Internet Stroke Center. URL http://www.strokecenter.org/patients/about-stroke/stroke-statistics/.

[2] WHO | The top 10 causes of death. *WHO*, 2017. URL http://www.who.int/mediacentre/factsheets/fs310/en/.

[3] Brian B Avants, Nick Tustison, and Gang Song. Advanced normalization tools (ants). *Insight j*, 2:1–35, 2009.

[4] Michael Eriksen, Majid Ezzati, Susan Holck, Carlene Lawes, Varsha Parag, Patricia Priest, and Stephen Vander Hoorn. The World Health Report. pages 14661–7000, 2002. URL http://www.who.int/whr/2002/en/whr02{_}en.pdf.

[5] Anders Hedblom. Blood vessel segmentation for neck and head computed tomography angiography, 2013.

[6] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank. Pet-ct image registration in the chest using free-form deformations. *IEEE Transactions on Medical Imaging*, 22(1):120–128, Jan 2003. ISSN 0278-0062. doi: 10.1109/TMI.2003.809072.

[7] D A Shrier, H Tanaka, Y Numaguchi, S Konno, U Patel, and D Shibata. Ct angiography in the evaluation of acute stroke. *American Journal of Neuroradiology*, 18(6):1011–1020, 1997. ISSN 0195-6108. URL http://www.ajnr.org/content/18/6/1011.

# Bio-Radar Model Validation using Chest-Wall Simulator

Carolina Gouveia [1]
Daniel Malafaia [1]
José Vieira [2]
Pedro Pinho [1]
Ana Tomé [2]
Pedro Magalhães [3]

[1] Instituto de Telecomunicações - 3810-193 Aveiro
[2] IEETA - Campus Universitário de Santiago 3810-193 Aveiro
[3] UA - Departamento de Eletrónica, Telecomunicações e Informática - Campus Universitário de Santiago 3810-193 Aveiro

## Abstract

The cardiopulmonary signal monitoring, without the usage of contact electrodes or any type of sensors, has several applications such as elderly's health monitoring, sleeping monitoring or even in search and rescue scenarios. The bio-radar system can measure vital signals accurately by using the Doppler effect principle, that relates the received signal properties with the distance variability between the radar antennas and the person's chest-wall. In this work, a mathematical model of bio-radar is presented. Furthermore it is proposed a fully controllable Chest-Wall Simulator (from now on called CWS) to validate the bio-radar system, as well as the proposed breathing extraction algorithm. Later results comprising simulated and real signals are discussed.

## 1 Introduction

The bio-radar's system is composed by a continuous wave Doppler radar which continuously transmits a sinusoidal carrier, generated digitally, and receives the echo from the reflecting target. Due to the Doppler effect, there is a phase change as the subject's chest-wall moves towards or away from the radar and hence a phase modulation in the received signal is created [1]. The overall system is represented by the block diagram in Fig.1, where the parasitic reflection $r_1(t)$ from nearby standing objects is also considered. Bio-signals have low amplitude and its bandwidth occupies very low frequency ranges close to DC, hence they are highly sensitive to several sources of noise, such as clutter from the scenario reflections. Therefore, in the next sections the mathematical model that synthesises the bio-radar system behaviour is presented. Also the CWS is described and used to evaluate the bio-radar's system and the real-time DSP algorithm developed for the bio-signals information extraction.

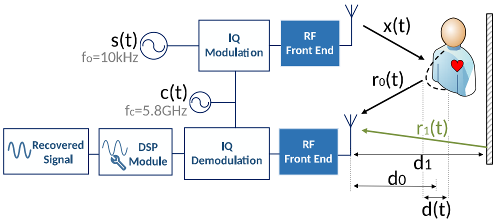


Figure 1: Bio-radar system block diagram.

## 2 Modelling the Bio-Radar system

### 2.1 Signal model for the Bio-Radar channel response

Regarding the block diagram presented in Fig.1, a RF signal $s(t) = e^{j\omega_0 t}$ is generated digitally, modulated with an in-phase and quadrature (IQ) modulation, with the carrier signal $c(t)$ and transmitted toward the target.

$$x(t) = \cos(\omega_0 + \omega_c)t \tag{1}$$

The received signal is expressed as:

$$\begin{aligned} r(t) &= r_0(t) + r_1(t) = \\ &= A_0 \cos((\omega_0 + \omega_c)t + \varphi(t)) + A_1 \cos((\omega_0 + \omega_c)t + \theta_1) \end{aligned} \tag{2}$$

where $A_0$ and $A_1$ are the amplitudes of the received signal from the subject and clutter, respectively. The signal $r_1(t)$ represents the sum of the total sources of clutter and its equation in the baseband is $r_1(t) = A_1 e^{j\theta_1} = A_1 e^{(j4\pi d_1/\lambda)}$. The phase change function which contains the respiratory information is represented by $\varphi(t)$. The chest-wall motion changes the wave's travelled distance and hence modulates the reflected signal. Thus, the phase change function can be described by the equation 3:

$$\varphi(t) = \theta_0 + \frac{4\pi d(t)}{\lambda} \tag{3}$$

where $\theta_0 = (4\pi d_o/\lambda) + \phi$ is the total path travelled by the wave, considering the nominal distance between the radar and the target, $d_o$, and the phase shift at the target's surface, $\phi$. In the remain of this work we consider the chest movement described as $d(t) = a_r \cos(2\pi f_1 t)$. For simulation purposes the chosen respiratory rate used was $f_1 = 0.3$Hz.

### 2.2 Breathing signal's extraction algorithm

The received signal is sampled by the RF front-end and IQ demodulated. This signal is then processed by a DSP algorithm represented by the diagram in Fig.2. The baseband complex signal $g(t)$ is downsampled once it is a narrowband low-pass signal.

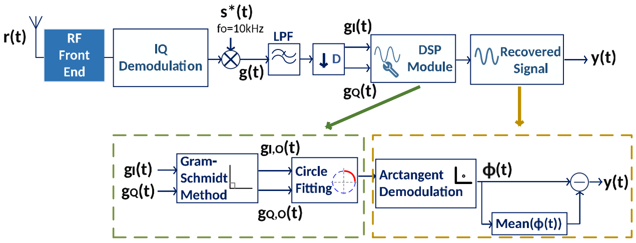
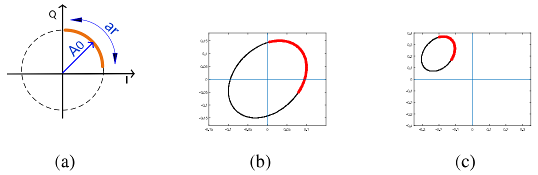


Figure 2: Receiver channel with DSP algorithm implementation.

The phase variation due to the target's motion, is represented in the polar plot by an arc (Fig.3(a)), where its length corresponds to the amplitude of the respiratory signal, $a_r$, and its radius is the received signal's amplitude $A_0$. In an ideal scenario the arc fits to a perfect circle centred in zero.



(a) (b) (c)

Figure 3: Complex plot of the baseband signal, $g(t)$, due to target's motion: (a) in an ideal scenario, (b) with IQ imbalance effect, (c) with DC offset and IQ imbalance effect.

However, in real-world scenarios there is IQ imbalance effect, which occurs when both real and imaginary parts do not have the same amplitude and the phase relationship is not exactly 90°. Hence, the formed arc fits to an ellipse instead a circle. There are also DC offsets present in both real and imaginary parts of the baseband signal, caused by the clutter, which leads to an offset in the arc's centre. Fig.3(c) shows the arc formed with the IQ imbalance effect and DC offset presence. These effects should be digitally removed before the phase demodulation in order to guarantee an accurate arctangent result.

The IQ imbalance can be removed by using the Gram-Schmidt method, by applying relation (4), [2], [1], which restores the orthogonality of the baseband signal in quadrature. The parameters $\psi_E$ and $A_e$ are phase and amplitude imbalance measured, respectively.

$$\begin{bmatrix} g_{I,O}(t) \\ g_{Q,O}(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\tan(\psi_E) & \frac{1}{A_e \cos(\psi_E)} \end{bmatrix} \begin{bmatrix} g_I(t) \\ g_Q(t) \end{bmatrix} \tag{4}$$

After the imbalance compensation, the DC offsets are estimated and removed using circle fitting method [3], which tracks the circle centre coordinates and subtract them from the complex signal, forcing the arc to be centred in zero. Finally, the arctangent is evaluated in order to extract the respiratory signal [1],[3], obtaining the signal $y(t)$, and its rate is computed using the power spectral density. In the simulation of this mathematical model using simulated signals was possible to recover the respiratory signal with rate of $f_1 = 0.3027$ Hz as expected.

## 3 Chest-Wall Simulator

The CWS replaces the human chest-wall in the diagram of Fig.1 and was developed in order to validate the mathematical model described previously and to test the DSP algorithm efficiency. This simulator is a mechanical platform that is pushed and pulled horizontally with a known motion rate which is, in this simulation, 0.4 Hz. To build this simulator was used a stepper motor, controlled by the Easydriver version 4.4 which controls the steps necessary to rotate the motor. The number of steps and the time between them is settled by a Arduino Pro Micro microcontroller. Fig.4 illustrates the main mechanical operations.

For model validation purposes, a few tests were preformed with two varying parameters and a total duration of 1 minute. Table 1 describes three testing scenarios, where the $a_r$ represents the amplitude of the chest-wall displacement $d(t)$, and $d_0$ is the distance between the radar and target that causes different position of the formed arc in the complex plot. The amount of chest-wall displacement in millimetres depends on the subject's body structure and should be between 4 and 12 mm, [2]. Tests are described as follows:

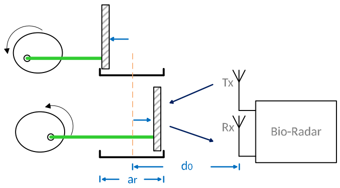


Figure 4: Mechanical schematic for the CWS motion.

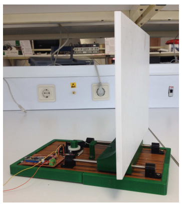


Figure 5: Chest-wall simulator.

**Test 1** – the CWS was used to analyse the impact of high amplitude variation, at a certain $d_0$ and a motion rate of 0.4 Hz;

**Test 2** – the CWS was used to analyse the impact of low amplitude variation, at a certain $d_0$ and a motion rate of 0.4 Hz;

**Test 3** – a subject was asked to synchronize his breathing with metronome clicks. The breathing rate was 24 breaths/ minute.

| | $a_r$ [cm] | $d_0$ [cm] |
|---|---|---|
| Test 1 | 0.7 | 56 |
| Test 2 | 0.3 | 66 |
| Test 3 | ≈ 0.4* | ≈ 70* |

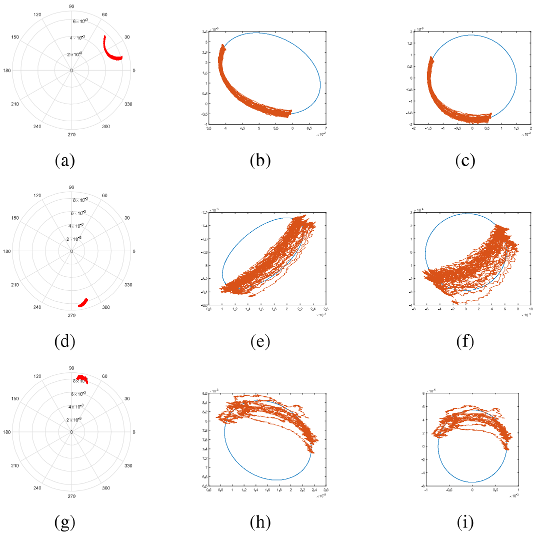
*Estimated values

Table 1: Parameters for each test. Tests 1 and 2 where performed with the CWS and the test 3 with a human.

## 4 Experimental Results

In this section, results of the three tests described previously will be shown and discussed. Polar plots before and after IQ imbalance compensation and DC offsets removal, of the performed tests are shown in Fig.6. In all the three testing scenarios, it was possible to do IQ imbalance compensation and DC offset removal successfully. It is possible to verify that for a larger target motion, it is easier to do an ellipse fitting and the imbalance compensation is more accurate, hence there are better results for the parameter $a_r = 7$mm. It is also possible to conclude that the parameter $d_0$ does not have any influence in the performance of the implemented algorithm. With the simulator, the major source of noise is the presence of friction while the mechanical platform moves forward and backward. The friction effect is more significant for lower displacement (when $a_r$ is 3mm).

Focusing now in test 3, when breathing it is impossible to keep the same breathing rhythm and it is difficult to stay completely stable during the acquisition. Comparing figures 6(b) and 6(c) with figures 6(h) and 6(i), it can be observed the following effect: the arc drawing is more disperse over the ellipse, in the case of the breathing signal than in the case of the simulator's first test.



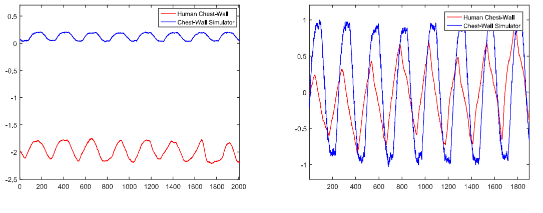
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 6: Baseband polar plot for test 1, 2 and 3: (a) (d) (g) after downsampling, (b) (e) (h) ellipse fitting for IQ imbalance compensation, (c) (f) (i) circle fitting after DC offsets removal and IQ compensation.

In order to clarify the proximity of the CWS extracted signal to a real breathing signal, was asked a subject to force his respiratory rate, as mentioned previously. Fig.7 shows the signals acquired as described in test 1 and test 3, after downsampling and after DC offsets removal. Note that figure 7(a) still has DC offset in both signals. Both waveforms are similar and could be equally extracted with the same rate of f1 = 0.3906 Hz.



(a) (b)

Figure 7: Arctangent result: (a) after downsampling, (b) after DC offsets removal and IQ compensation.

## 5 Conclusion

In this work a mathematical model of the bio-radar system was introduced. A DSP algorithm for phase demodulation was presented and evaluated using a chest-wall simulator built for this purpose. This simulator was projected to move at a constant rate of 0.4 Hz, in order to validate the results of the extracted respiratory signal. By applying the signals captured with the simulator in the algorithm, it was possible to verify that the algorithm works regardless of the position of the target or its chest-wall motion. The same algorithm was tested using a real acquired respiratory signal and the same conclusions were achieved.

## References

[1] Olga Boric-Lubecke, Victor M Lubecke, Amy D Droitcour, Byung-Kwon Park, and Aditya Singh. *Doppler Radar Physiological Sensing*. John Wiley & Sons, 2015.

[2] Amy Diane Droitcour et al. *Non-contact measurement of heart and respiration rates with a single-chip microwave doppler radar*. PhD thesis, Stanford University, 2006.

[3] Byung-Kwon Park, Victor Lubecke, Olga Boric-Lubecke, and Anders Host-Madsen. Center tracking quadrature demodulation for a doppler radar motion detector. In *Microwave Symposium, 2007. IEEE/MTT-S International*, pages 1323–1326. IEEE, 2007.

# Tracking Large Anterior Mitral Leaflet Displacements by Incorporating Optical Flow in an Active Contours Framework

Malik Saad Sultan
msultan@dcc.fc.up.pt
Miguel Tavares Coimbra
mcoimbra@dcc.fc.up.pt

Instituto de Telecomunicações,
Faculdade de Ciências da Universidade do Porto,
Porto, Portugal

## Abstract

Echocardiography is an important tool to detect early evidence of mitral valve degradation associated with rheumatic heart disease. The segmentation and tracking of the Anterior Mitral Leaflet helps to quantify the morphologic valve anomalies, such as the leaflet thickening, shape and the mobility changes. The tracking of this leaflet throughout the cardiac cycle is still an open challenge in the research community. The widely used active contours segmentation framework fails when faced with large leaflet displacement. In this work, we propose the integration of optical flow in an open-ended active contour framework to address this difficulty. This additional information promotes solutions with contours next to high leaflet displacements, resulting in superior performance. The algorithm was tested on 9 fully annotated real clinical videos, acquired from the parasternal long axis view. The algorithm is compared with our previous work. Results show a clear improvement in situations where the leaflet exhibits large displacement or irregular shapes, with an average error of 4.5 pixels and a standard deviation of 2 pixels.

## 1 Introduction

Rheumatic heart disease (*RHD*), among other causes, can be a serious consequence of repeated episodes of acute rheumatic fever (*ARF*). *RHD* is still a major problem in developing countries, affecting about 15 million people worldwide. The disease is responsible for 233,000 deaths per year, with 282,000 new registered cases [1].

The mitral valve is the most commonly affected heart valve by *RHD*, with or without the involvement of the other valves [2]. The parasternal long axis (*PLAX*) is the best window to analyze the mitral valve (Fig. 1) [3]. To measure and quantify thickness, shape and the motion pattern of the mitral leaflets, we need a robust segmentation and tracking framework that can automatically delineate the anatomical structures in the image.

The active contour framework [4] is widely used by the research community to segment and track the structures in medical images/videos. They are the parametric contours that deform under the image characteristics and other shape constraints.

This method has several limitations such as high dependency on the initial contour placement or high sensitivity to edge and noise information. The research community has addressed some of these problems, by introducing improvements such as pressure force (balloon) [5] and gradient vector flow [6].

The segmentation and tracking of the mitral leaflets in ultrasound was addressed in previous works [7 - 9]. Based on the fact that the motion of the cardiac muscles are different from the mitral leaflet, two active contours segments were used and are initialized by the rough segmentation obtained through the curve fitting algorithm [7]. Optical flow is used to obtain the initial boundaries that are then refined by the active contours [8]. The motion of the mitral leaflet is very irregular, it rotates, translates and shows large leaflet displacement and thus proposed approach fails in those situations. The internal and external energy of the classical snake model is modified to obtain robust tracking of the anterior mitral leaflet (*AML*) [9]. Open-ended active contours were used and the external energy of the model was adapted to encourage the end point of the contour to continually follow the tip of the *AML*. However, the proposed approach fails in situations when the *AML* shows large displacement.

In this work, our main objective is to improve the performance of the segmentation and tracking algorithm by addressing situations with large displacements. As contributions, our external energy is modified to incorporate optical flow energy, promoting solutions associated with the large motion exhibited by the mitral valve leaflet. This work builds upon our previous research [9] that frequently encounters failure in such situations.

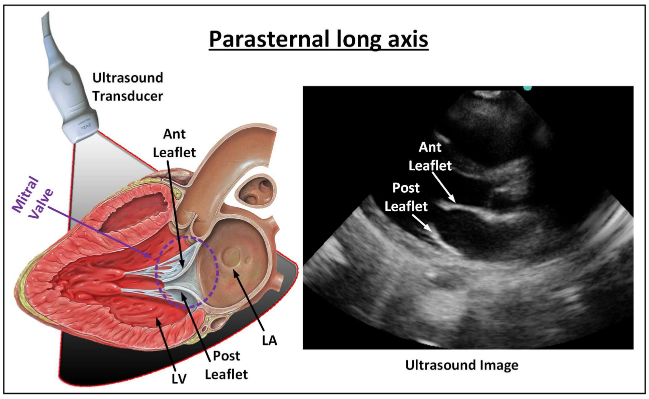


Figure 1: Parasternal long axis view (adapted from [9]) of a normal mitral valve in diastole.

This paper is organized as follows. Section 2 provides the overview of the classical snake model. Section 3 presents the details of improved snake model. Finally, section 4 reports the tracking results of the *AML*.

## 2 Classical Snake

The active contour model, also known as snakes, uses parametric curves $V(S) = (x(S), y(S))^T$ $S \in [0,1]$ that deform under the image force, limited by shape constraints. The total energy of the snake model consist of a combination of internal and external energy(1).

$$\int_0^1 \left( \overbrace{\frac{1}{2}\left[ \underbrace{\alpha \left\| \frac{\partial V}{\partial S} \right\|^2}_{Elasticity} + \underbrace{\beta \left\| \frac{\partial^2 V}{\partial S^2} \right\|^2}_{Stiffness} \right]}^{Internal} + \overbrace{E_{ext}(V)}^{External} \right) ds \tag{1}$$

## 3 Improved Snake Model

The focus of this work is to improve the tracking capabilities of our previous algorithm [9]. We assume perfect segmentation in the very first frame, which is then tracked throughout the cardiac cycle. Since this is an extension of our previous work, the integration of the optical flow in external energy of the classical snake model is here explained in more detail.

### 3.1 External energy

The external energy is divided into two main parts, the image and the motion energy (2).

$$E_{ext.} = E_{image} + E_{motion} \tag{2}$$

The image energy is the same used in our previous work [9]. The motion energy quantifies the image displacement energy for overcoming the limitation (tracking failure in large motion) of our previous work. Optical flow is integrated in the external energy of the classical snake model as a dynamic energy (3).

$$E_{motion} = w_{optical\_flow} E_{optical\_flow} \tag{3}$$

We selected the Lucas-Kanade approach, considered by literature as still the most interesting cost-benefit optical flow method [10].

In this work, we used a coarse-to-fine approach to Lucas-Kanade to efficiently handle both large and small motion. This consists on creating multiple copies of the same image with different resolutions. In each level, the resolution is reduced $1/4$ of the size of the previous level.

At this point, we have the motion vectors for each pixels that give us the displacement in relation to the previous frame. Our objective is to integrate this information (displacement and direction) in the classical snake model. We added the motion vectors as the displacement force in an external energy that pushes the contour points towards the true image energy, until trapped in local minimum. In simple words, the optical flow is used as a displacement force that pushes the curve towards the local minimum.

The iterative form of the modified classical snake model is shown (4, 5)

$$E_{ext.} = \begin{bmatrix} \frac{\partial E_{image}}{\partial x} \\ \frac{\partial E_{image}}{\partial y} \end{bmatrix} + \begin{bmatrix} \frac{\partial E_{motion}}{\partial u} \\ \frac{\partial E_{motion}}{\partial v} \end{bmatrix} = \begin{bmatrix} f_x \\ f_y \end{bmatrix} + \begin{bmatrix} f_u \\ f_v \end{bmatrix} \tag{4}$$

$$\begin{aligned} X_t = &\quad (A+\gamma I)^{-1}(\gamma X_{t-1} - f_x(X_{t-1}, Y_{t-1}) \\ &\qquad\qquad + f_u(X_{t-1}, Y_{t-1})) \\ Y_t = &\quad (A+\gamma I)^{-1}(\gamma Y_{t-1} - f_y(X_{t-1}, Y_{t-1}) \\ &\qquad\qquad + f_v(X_{t-1}, Y_{t-1})) \end{aligned} \tag{5}$$

Whereas, $I$ is the identity matrix, $\gamma$ is the step size and $A$ is coefficient matrix. The first part of the equation $(A+\gamma I)^{-1}$ imposes the internal constraints and the other part attracts the contour toward the image energy.

## 4 Results

### 4.1 Materials

In one of the activities of Real Hospital Português, in Recife, Brazil, a dataset of ultrasound mitral valve videos (*PLAX* view) has been collected for the purposes of screening acute rheumatic fever in children. The data was collected using a M-Turbo model by SonoSite ultrasound system, with $C11x$ transducer (5-8 MHz). Nine of these exams were fully annotated (manual segmentation of all frames) using support software and were used to test the novel algorithm proposed in this work. These nine videos include a total of 1137 frames with dimensions of $[351 \times 441]$.

### 4.2 Tracking results

The tracking starts with the manual initialization by inserting 6 points on the *AML*, in the first frame of each video. To evaluate the performance, results of the reference approach [9] and the proposed approach are compared with the ground truth (doctor's annotation). The Modified Hausdorff Distance (*MHD*) [11] error is computed for the ground truth and the automatic segmentation. Table 1 shows the mean, median and the standard deviation of the *MHD* error in each video, for both the reference and our approach. The results are computed for the most difficult situation, when the *AML* shows large displacement, for example, at first opening of the *AML* and in the diastole phase. The results shows improvement in each video, with the overall mean, median and standard deviation of 2, 4.5 and 4.2 pixels, respectively.

Better tracking performance is achieved but at the cost of computational complexity. The reference algorithm [9] consumes about 0.58 second/frame to automatically delineate the *AML*. However, our approach takes about 14.07 second/frame.

Table 1: TRACKING RESULTS (in pixels)

| Video No. | Our Approach<br>STD / Mean / Median | Ref. Approach [9]<br>STD / Mean / Median |
|---|---|---|
| **1** | 2.2 / 3.2 / 2.5 | 3.2 / 6.0 / 5.8 |
| **2** | 2.5 / 6.0 / 5.5 | 4.0 / 9.9 / 11.3 |
| **3** | 1.7 / 4.2 / 3.9 | 2.5 / 5.8 / 5.3 |
| **4** | 2.2 / 4.9 / 4.5 | 1.4 / 5.6 / 5.1 |
| **5** | 1.6 / 4.5 / 4.6 | 2.7 / 6.4 / 5.7 |
| **6** | 3.4 / 5.3 / 4.9 | 6.2 / 7.9 / 5.3 |
| **7** | 0.7 / 4.7 / 4.6 | 1.1 / 5.1 / 4.7 |
| **8** | 0.9 / 3.5 / 3.6 | 0.9 / 4.8 / 4.8 |
| **9** | 3.2 / 4.5 / 3.7 | 3.5 / 5.4 / 4.2 |
| **Total** | **2.0 / 4.5 / 4.2** | **2.8 / 6.3 / 5.8** |

## Acknowledgement

This article is a result and funded by the project NanoSTIMA, NORTE-01-0145-FEDER-000016, supported by Norte Portugal Regional Operational Programme (NORTE 2020), through Portugal 2020 and the European Regional Development Fund (ERDF) and project UID / EEA / 50008 / 2013. It is also a result of the project (NORTE-01-0247-FEDER-003507-RHDecho), co-funded by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF) and Fundação para a Ciência a e Tecnologia (FCT) grant no: PD / BD / 105761 / 2014.

## References

[1] Carapetis JR, Steer AC, Mulholland EK, Weber M. The global burden of group A streptococcal diseases. Lancet Infect Dis. 2005;5:685âĂŞ694

[2] B. Remnyi, N. Wilson, A. Steer, B. Ferreira, J. Kado, K. Kumar, J. Lawrenson, G. Maguire, E. Marijon et. al. Feb 2012. World Heart Federation criteria for echocardiographic diagnosis of rheumatic heart diseasean evidence-based guideline, Nat Rev Cardiol. Vol 9, issue 5, pp 297-309

[3] A.S. Omran, A.A. Arifi, A.A. Mohamed, Echocardiography of the mitral valve, Journal of the Saudi Heart Association, vol. 22, no. 3, pp. 165170, Feb. 2010

[4] Kass M., Witkin A., and Terzopoulos D., âĂIJSnakes: Active contour models,âĂİ Int. J. Comput. Vis. 1, 321âĂŞ331 (1988).

[5] L. D. Cohen, On active contour models and balloons, CVGIP: Image Understanding, vol. 53, no. 2, pp. 211218, 1991

[6] C. Xu and J. Prince. Gradient Vector Flow: A new External Force for Snakes. In CVPR, pages 66âĂŞ71, Puerto Rico, USA, 1997

[7] S. Martin, V. Daanen, O. Chavanon, J. Troccaz. Fast Segmentation of the Mitral Valve Leaflet in Echocardiography. Computer Vision Approaches to Medical Image Analysis, Vol. 4241, pp 225-235, 2006

[8] I. Mikic, S. Krucinski, J. D. Thomas. Segmentation and Tracking in Echocardiographic Sequences: Active Contours Guided by Optical Flow Estimates. IEEE transactions on medical imaging, vol- 17, no. 2, April 1998

[9] M. S. Sultan, N. Martins, D. Veiga, M. J. Ferreira, M. T. Coimbra, Tracking of the Anterior Mitral Leaflet in Echocardiographic Sequences using Active Contours, EMBC, 1074-1077, 2016

[10] J.L. Barron, D.J. Fleet, S.S. Beauchemin, Systems and Experiment Performance of Optical Flow Techniques, International Journal of Computer Vision, 12:1, 43-77 (1994)

[11] M.-P. Dubuisson, A. K. Jain, 1994. A Modified Hausdorff Distance for Object Matching, Proc, international conference on pattern recognition, Jerusalem, Israel, pp 566-568

# Information discriminant analysis for improved feature extraction in heart sound segmentation

Francesco Renna
fraренna@dcc.fc.up.pt
Jorge Oliveira
oliveira_jorge@dcc.fc.up.pt
Miguel T. Coimbra
mcoimbra@dcc.fc.up.pt

Instituto de Telecomunicações
Faculdade de Ciências da Universidade do Porto
Porto, Portugal

## Abstract

In this work, we present a method to extract features from heart sound signals that enhances segmentation performance. The way features are extracted from the recorded signals is adapted to the data itself. The proposed method is based on the extraction of delay vectors, which are modeled with Gaussian mixture model priors, and an information-theoretic dimensionality reduction step which aims to maximize discrimination between delay vectors in different segments of the heart sound signal.

We report preliminary experimental results where we test our approach with heart sounds from the DigiScope dataset showing an average positive predictive value per samples of 77.4%.

## 1 Introduction

Auscultation is one of the fundamental steps of the physical examination of a patient and the first line of screening of cardiovascular disease. Based on the clinical signs extracted from it, a clinician decides if additional exams are needed, typically an echocardiogram, which is more reliable, but also more expensive and requiring specific expertise to apply. The cost-effectiveness of auscultation contrasts with the worrying revelations of important studies quantifying that only around 20% of medical interns can actually perform cardiac auscultation effectively [5]. This observation fueled recent research efforts in automatizing part or the entire process of analysis of the phonocardiogram (PCG) signal.

A key step usually implemented to extract information from a heart sound signal consists in dividing each heart cycle into its fundamental components. Namely, a normal heart cycle is usually divided into first heart sound (denoted by $S_1$), systole, second heart sound (denoted by $S_2$) and diastole. Further sound components of interest are represented by murmurs, clicks, splitting of the first and/or second heart sounds and further third ($S_3$) and fourth ($S_4$) heart sounds.

Different approaches have been presented in the literature in order to perform PCG segmentation (see [4] for a general overview). In general, such methods are performed via a two-steps approach, that first extract features from the PCG signal and then assigns features to the different states corresponding to the different sound segments or components. In particular, features can be extracted in the time domain (e.g., Shannon energy), in the frequency domain (e.g., Mel-frequency cepstral coefficients) or other transform domains (wavelet transform, etc.).

Then, features corresponding to different time instants can be assigned to the different segments of the signal using various kinds of classifiers (e.g., support vector machines (SVM) [9], neural networks [1], etc.).

Another popular approach is based on the explict modeling of the sequential nature of states in the heart sound signal. In particular, algorithms based on hidden Markov models (HMM) are designed in order to segment heart sound signals. The use of HMM for heart sounds dates back to the work of Gill *et al.* [2] and its evolutions have been recently shown to achieve state-of-the-art segmentation performance [8].

However, PCG segmentation still represents a challenging problem to solve when considering its application in real-world, noisy environments. In this work, we propose a technique to extract features from the heart sound signal, which is adapted to the data via training, and which aims to increase separability between segments corresponding to different signal states. The method provides a principled framework based on dynamical system analysis (delay vectors method [3]) and information-theoretic tools to enhance segmentation performance. In particular, the proposed algorithm is based on the following steps: i) extract *delay vectors* from a heart sound signal; ii) model the delay vectors with state-dependent, multivariate Gaussian mixture model (GMM) priors; iii) reduce the delay vector state space by using information-theoretic linear discriminant methods.

The proposed feature extraction technique can be coupled with different kinds of classifiers. In particular, it can be applied to HMMs due to the amenability in computing the statistical quantities required for optimal estimation of the hidden state sequence.

## 2 Methods

### 2.1 Delay vectors

We denote with $x(t), t = 0, 1, \dots, T$ the samples of the observed heart sound signal, and with $s(t), t = 0, 1, \dots, T$ the corresponding labels, i.e., $s(t) \in \{1, \dots, I\}$, where the values $1, \dots, I$ correspond to different states of the PCG signal. In particular, the PCG states considered in this work are: $S_1$, systole, $S_2$ and diastole.

In dynamical system analysis, delay vectors are used to study the evolution of nonlinear systems, and they are shown to contain the full information about the system dynamics under mild assumptions [3]. Delay vectors are extracted from the observed signal as follows: given a time lag $\tau \in \mathbb{N}$ and a vector dimension $n \in \mathbb{N}$, the delay vector associated to time $t$ is defined as $\mathbf{x}(t) = [x(t), x(t-\tau), \dots, x(t-(n-1)\tau)]^{\mathrm{T}}$, where $(\cdot)^{\mathrm{T}}$ denotes the transpose operator.

Then, delay vectors corresponding to a given state $i$ are modeled via a multivariate GMM, i.e., we assume that the probability density function (pdf) of the vector $\mathbf{x}(t)$ conditioned on $s(t) = i$ is given by

$$b_i(\mathbf{x}(t)) = p(\mathbf{x}(t)|s(t) = i) = \sum_{k=1}^{K^{(i)}} p_k^{(i)} \mathcal{N}(\mu_k^{(i)}, \Sigma_k^{(i)}), \tag{1}$$

where $K^{(i)}$ is the number of Gaussian components in the GMM emission distribution associated to state $i$ and $p_k^{(i)}$ is the probability of the $k$-th component in state $i$, so that $0 \le p_k^{(i)} \le 1$ and $\sum_k p_k^{(i)} = 1$. Moreover, $\mathcal{N}(\mu_k^{(i)}, \Sigma_k^{(i)})$ denotes a Gaussian multivariate distribution with mean $\mu_k^{(i)}$ and covariance matrix $\Sigma_k^{(i)}$.

Note that the statistical description of a delay vector $\mathbf{x}(t)$ conditioned on state $i$ does not depend on the time instant $t$. For this reason, we can drop the time index and simply write $b_i(\mathbf{x}) = p(\mathbf{x}|s = i)$. Moreover, the marginal distribution of a delay vector $\mathbf{x}$ is obtained by averaging the conditional distributions as

$$p(\mathbf{x}) = \sum_{i=1}^{I} \pi_i \sum_{k=1}^{K^{(i)}} p_k^{(i)} \mathcal{N}(\mu_k^{(i)}, \Sigma_k^{(i)}), \tag{2}$$

where $\pi_i$ denotes the prior probability of being in state $i$.

Finally, we assume that the parameters $\pi_i, p_k^{(i)}, \mu_k^{(i)}, \Sigma_k^{(i)}$, $i = 1, \dots, I$, that define the statistical description of the delay vectors, are estimated from annotated data in the training phase. Then, the distributions $p(\mathbf{x})$ and $p(\mathbf{x}|s = i)$ are used in the next step of the proposed feature extraction process, that is, dimensionality reduction.

### 2.2 Information-theoretic dimensionality reduction

Dimensionality reduction consists in finding a linear transformation which aims to: i) reduce the dimension of the delay vectors in order to decrease computational complexity and ii) to enhance discrimination between delay vectors belonging to different states.

In particular, when $m \in \mathbb{N}$ features are extracted from each delay vector, on denoting by $\Phi \in \mathbb{R}^{m \times n}$ the linear transformation applied to the delay vectors, we quantify the level of discrimination between transformed delay vectors in different states by using an approximation of Shannon's mutual information, i.e., we write

$$\mathbb{I}_G(\Phi\mathbf{x};s) = h_G(\Phi\mathbf{x}) - \sum_{i=1}^{I} \pi_i h_G(\Phi\mathbf{x}|s=i), \tag{3}$$

where $h_G(\mathbf{x})$ is the differential entropy of a Gaussian vector having the same mean and covariance matrix as the vector $\mathbf{x}$. Then, the desired $\Phi$ is obtained as the solution of the optimization problem

$$\underset{\Phi}{\text{maximize}}\ \mathbb{I}_G(\Phi\mathbf{x};s), \quad \text{subject to } \mathrm{tr}(\Phi\Phi^{\mathrm{T}}) \leq m, \tag{4}$$

where $\mathrm{tr}(\cdot)$ is the trace operator and the constraint in (4) is imposed in order to guarantee stability with respect to measurement noise. This approach that extract features from data aiming at maximizing discrimination via an approximation of Shannon's mutual information goes under the name of information discriminant analysis (IDA) [6]. It is also possible to show that $\mathbb{I}_G(\Phi\mathbf{x};s)$ is a concave function of $\Phi$ and its gradient with respect to $\Phi$ can be expressed in closed form. In this way, convergence to a global optimum can be guaranteed when using gradient descent [6].

### 2.3 Proposed features with HMMs

The proposed method for feature extraction can be easily incorporated in algorithms for heart sound segmentation which are based on HMMs. In particular, we recall that a HMM is completely defined by a set of initial state probabilities, state transition probabilities and emission probability distributions. Then, on denoting by $\mathbf{y}(t) = \Phi\mathbf{x}(t), t = 0,1,\ldots,T$ the sequence of observed, transformed delay vectors, these represent the observed emissions in the HMM. Moreover, the corresponding emission pdfs can be easily computed from (1) as

$$b_i(\mathbf{y}) = p(\mathbf{y}|s=i) = \sum_{k=1}^{K^{(i)}} p_k^{(i)} \mathcal{N}(\Phi\mu_k^{(i)}, \Phi\Sigma_k^{(i)}\Phi^{\mathrm{T}}). \tag{5}$$

Finally, the state sequence that maximizes the log-likelihood associated with the sequence of observed feature vectors $\mathbf{y}(t)$ can be efficiently computed via the use of the Viterbi algorithm.

## 3 Experimental methodology

In this section, we present an application example of the proposed feature extraction framework for PCG segmentation. In particular we use heart sounds from the DigiScope dataset. Such dataset is composed of samples from 29 different healthy individuals, ranging in age from six months to 17 years old. The recordings have a minimum, maximum and average duration of $\approx 2, 20$ and 8 seconds, respectively. Heart sounds were collected in Real Hospital Português (Recife, Brasil) using a Littmann 3200 stethoscope embedded with the DigiScope Collector [7] technology, recorded at 4000 Hz. The heart sounds have all been collected from the mitral area. Then, these sounds were manually annotated by cardiopulmonologists using the audacity software[1].

Experiments are conducted as follows: the considered 29 heart sound signals are splitted randomly so that 14 sounds are used for training and 15 sounds are used for testing. Delay vectors are extracted from both training and testing raw signals by choosing $\tau = 100$ and $n = 32$, so that each delay vector spans approximately 0.77s of the recorded PCG signal. In this sense, we have observed experimentally that the best segmentation performance is achieved when allowing the delay vectors to span a large portion of the PCG signal, approximately an entire heart beat.

Annotated training data are used to determine the delay vector distributions $b_i(\mathbf{x})$. In particular, we choose the number of Gaussian components in each state-dependent GMM as $K^{(i)} = 8$, $i = 1,\ldots,I$ and the parameters $p_k^{(i)}, \mu_k^{(i)}$ and $\Sigma_k^{(i)}$ are estimated for each state $i$ by using the expectation maximization (EM) algorithm. Then, we fix $m = 20$ and we compute the matrix $\Phi \in \mathbb{R}^{20\times 32}$ by following the steps described in Section 2.2. Such $\Phi$ is applied to the delay vectors extracted from the sounds in the testing set, and the emission probability distributions are obtained as in (5).

Initial state probabilities and state transition probabilities are estimated via standard maximum likelihood estimators.

Finally, the performance of the proposed segmentation method is measured in terms of the positive predictive value per sample, which accounts for the general capacity of the algorithm in discriminating sounds belonging to $S_1$, systole, $S_2$ and diastole intervals.

### 3.1 Preliminary results

The positive predictive value is computed on a patient basis, and then averaged over the testing set. Then, the means and standard deviation of these values over 100 different random training/testing splits are considered, for both cases when no linear dimensionality reduction is applied to the delay vectors, and for the case implementing the IDA approach described in Section 2.2. In particular, no linear dimensionality reduction guarantees a positive predictive value per sample of $71.1\% \pm 16.3\%$, whereas the IDA approach reaches a positive predictive value per semple of $77.4\% \pm 3.2\%$.

In this sense, we observe that information-theoretic linear dimensionality reduction plays an important role in enhancing discrimination between delay vectors of different signal states, thus leading to better segmentation performance. Moreover, it also reduces the variations between the outputs of different train/testing splits.

## 4 Conclusion

In this work, we have proposed a method to extract features from the PCG signal to enhance segmentation performance. The method, which is based on the extraction of delay vectors and on information discriminant analysis, is shown to produce promising results when applied to the DigiScope heart sounds dataset.

## Acknowledgements

This work is a result and funded by the project NanoSTIMA (NORTE-01-0145-FEDER-000016), supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF). It is also a result of the internal project SmartHeart in scope of project UID/EEA/50008/2013.

## References

[1] T.-E. Chen et al. S1 and S2 heart sound recognition using deep neural networks. *IEEE Trans. Biomed. Eng.*, 64(2):372–380, 2017.

[2] D. Gill, N. Gavrieli, and N. Intrator. Detection and identification of heart sounds using homomorphic envelogram and self-organizing probabilistic model. In *Computers in Cardiology*, pages 957–960, 2005.

[3] H. Kantz and T. Schreiber. *Nonlinear time series analysis*, volume 7. Cambridge university press, 2004.

[4] C. Liu et al. An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37(12):2181–2213, 2016.

[5] S. Mangione. Cardiac auscultatory skills of physicians-in-training: a comparison of three english-speaking countries. *The American journal of medicine*, 110(3):210–216, 2001.

[6] Z. Nenadic. Information discriminant analysis: Feature extraction with an information-theoretic objective. *IEEE Trans. Patt. Anal. Mach. Intell.*, 29(8):1394–1407, Aug. 2007.

[7] D. Pereira et al. DigiScope - Unobtrusive collection and annotating of auscultations in real hospital environments. In *IEEE EMBC*, pages 1193–1196, 2011.

[8] D. B. Springer, L. Tarassenko, and G. D. Clifford. Logistic regression-HSMM-based heart sound segmentation. *IEEE Trans. Biomed.l Eng.*, 63(4):822–832, 2016.

[9] J. Vepa. Classification of heart murmurs using cepstral features and support vector machines. In *IEEE EMBC*, pages 2539–2542, 2009.

[1] www.audacityteam.org

# Segmentation of Spacial Aliasing Artefact on MRI

João F. Teixeira[1]
jpft@inescporto.pt

Hélder P. Oliveira[1,2]
hfpo@inesctec.pt

[1]INESC TEC
Porto, PT

[2]Faculty of Sciences, University of Porto
Porto, PT

## Abstract

Magnetic Resonance Imaging (MRI) exams suffer from undesirable structure replicating and overlapping effects on certain acquisition settings. These are called Spatial Aliasing Artefacts (SAA) and their presence interferes with the segmentation of other anatomical structures.

This paper addresses the segmentation of the SAA, in order to remove their influence. The proposed method comprises thresholding, Region Growing, a contour refinement using Convex Hull and a Minimum Path algorithm applied to two orthogonal planes (Sagittal and Axial). Experiments on this processing pipeline report promising results.

## 1 Introduction

The work developed concerns the identification of undesirable markings on MRI sets (Figure 1(a)) and is based on [8]. The purpose of such delineation is to remove them, or mitigate their effects, preventing from interfering with the segmentation of other anatomical structures. These specific marking can be found on the centre of patient's body, and do not correspond to any expected anatomical structure for that position. These SAA, are a known consequence of having body regions outside the machine's field of view, during a MRI exam [3]. In this work, this effect concerns the shifting of cropped copies of shoulders to both centre and opposing side of the body.

Structure segmentation is a prevalent goal for medical image processing tasks. Similar works on MRI images, in particular concerning breast cancer, are not that frequent, with the exception of lesion detection aiming for diagnosis. In fact, the focus of the detection methods is on the breast area rather than the whole body, and thus are not affected by the SAA. To the best of our knowledge there has been no published work trying to identify or remove this kind of interference. Thus, these experiments had to rely on general purpose methods for segmentation.

Due to the images' settings, the obvious first segmentation choice could be a combination of adaptive thresholding, such as *Otsu's method* [4], and connected components selection based on their properties, such as shape and location. However, as Figure 1(a) suggests, a unique thresholding approach would be insufficient since some cases' shape include an outward portion with uneven intensity.

*Region Growing* [4] has the potential to iteratively aggregate regions that may not follow the majority of a distribution. This, however, also has the potential to include a considerable amount of inaccurate regions if the inclusive criteria are not strict enough. Thus, the balance of the method's configuration is delicate.

Another promising method for the segmentation of roughly regular shapes is *Active Contours* [1], or Snakes. These require an energy map, usually gradients, in order to enhance edges of objects, to which the contour progresses iteratively. Similarly to *Region Growing*, the *Snakes*' results may vary significantly depending on the input map provided. In the case depicted in Figure 1(a), the gradient may induce inward dents to the contour, due to the intensity pattern disparity of the region. On the other hand, *Active Contours* also needs to configure elasticity and curvature of the contours, not a trivial task concerning the anatomical variability.

The results of a method, downstream in the pipeline, may be significantly changed by a pre-processing step. Image enhancement has a broad array of available techniques that could be applied. One of which is intensity expansion/compression through manipulation of the images' gamma value. This non-linearity alters the influence of intensity ranges. Provided the objects' dynamic range includes other structures, this approach would most likely provide minimal upturn. Denoising through Wavelets decomposition [4] is also a common enhancement approach. Regional denoising methods such as mean and median filters [4] have been largely used to remove background noise. Shannon Entropy [9] and Homogeneity [6] measures have also been successful on segmentation tasks.

Despite the lack of a particular methodology to apply to this problem, several generic approaches seem likely to produce satisfactory results. In this work, some of the most promising and least subject to parameter fine-tuning were combined. The focus of the method concerns restricted path optimization to obtain accurate contour of objects with fairly consistent shape, in spite of sometimes having uneven intensity distributions.

## 2 Methodology

This work focuses on delineating the SAA's central copies due to their particular brightness. The outcome could then be shifted and fitted to the remaining lateral copies of the SAA. The proposed solution combines four modules: the first comprises thresholding and object selection. Second, region growing is applied over an entropy map. Then, to approximate the contours to their correct shape (for objects with gradual intensity decay), a convex hull is used. To obtain the precise SAA contours, a *Minimum Path* algorithm is used across the slices. Due to some misalignment of the left/right shoulder copies, each half was processed independently.

### 2.1 Base Processing Modules:

On a histogram, the desired intensity region generally presents a much broader and lower peak than the background image. As such, the base segmentation employs the *Triangle thresholding method* [10], designed to solve similar cases. Afterwards, the objects outside the sagittal and coronal profiles (x and y axis, respectively) are removed, on the basis that the artefacts occupy a large enough number of voxels.

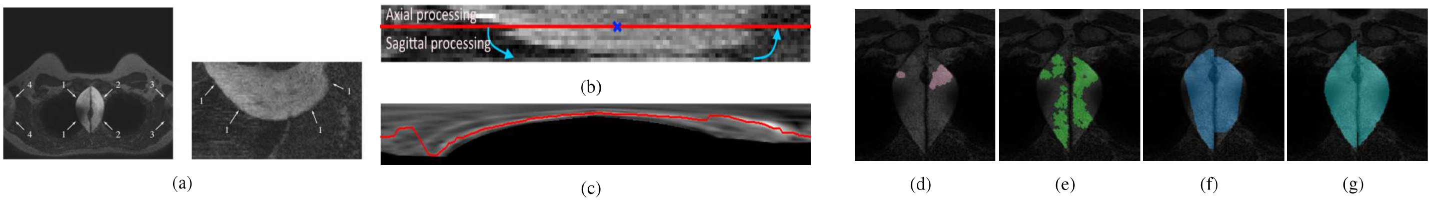
Extending the previous segmentation, *Region Growing* [4] was applied to the volume. This method requires three inputs: a 3D map ($\beta$), seed points and inclusion criteria. An adaptation of *Shannon Entropy* measure [9] produces interesting results, concerning the diminishing of the effect of noise. The seed points are based on the previous segmentation. Each axial object was lightly eroded and thinned, generating new contours, that were sampled (1/4 factor). These 3D contour samples were used as seed points, preserving the shape of the previous segmentation. Lastly, connectivity 8 was used to include points within a range of $2.5\sigma$ of the $\mu$ intensity at the seeds' positions.

In some cases (Figure 1(a)) the *Region Growing* is insufficient to close the degraded part of the SAA. As artefacts refer to shoulder anatomy, they are mostly convex, both in axial slices and across the sagittal axis, from bottom to top. Thus, gaps were compensated using *Convex Hull* (*QHull*).

### 2.2 Bi-perspective Polar Minimum Path:

This module constitutes a particular adaptation of the work of Oliveira *et al.* [5]. That work consisted in obtaining the contour using a *Polar Minimum Path* algorithm that is restricted, only moving forward or in the forward diagonal, from margin to margin, turning the image into a graph. Similarly here, each half (left and right), axial and sagittal slice-wise, consists of quasi-elliptic shapes. Hence, the Cartesian to Polar conversion was restricted to half the angle span. A transformation of a sagittal slice is depicted in Figure 1(b), with the *Minimum Path* result in Figure 1(c).

The *Polar Minimum Path* behaves poorly in the presence of small objects. Conversely, the shape of the SAA is uneven, since the bottom region appears larger on the sagittal plane and, similarly, the top region on the axial plane. This setting favors a bi-parted application of the *Polar Minimum Path*, starting on the sagittal plane and then moving to the axial plane for processing the remaining top slices. The centre positions for the *Polar Minimum Path* are found using the centroid of the previous segmentation. The Sagittal/Axial processing interface slice is the middle slice containing object from the previous segmentation.



(a) (b) (c) (d) (e) (f) (g)

Figure 1: (a) MRI views with SAA locations, (b) Sagittal slice with processing interface and Polar direction, (c) Polar transformed Sobel of (b) with *Minimum Path* (red), (d-g) Processing results: (d) Thresholding, (e) Region Growing, (f) Conex Hull, (g) Final *Polar Minimum Path*.

The two perspectives also present undesirable objects in a slightly different fashion, hence the *Minimum Path* cost maps were calculated differently. For the bottom part a Sobel vertical gradient sufficed. The map for the top part was enhanced by a 75% Top-hap operation of the normalized image, and 25% of the complementary of the result of Laplacian of Gaussian filter on the original image. The values were empirically tuned to reduce noise and enhance high gradients.

Moreover, the image input of the *Minimum Path* is further processed prior to being used. To prevent misinterpretation of undesired high gradient regions over desired low gradient ones, regions are blocked by a mask calculated from the previous module's segmentation. The mask is produced by eroding and dilating the existing objects and subtracting them to produce a curved strip with a shape similar to the desired path. The masked-out region is shown as black in Figure 1(c).

## 3 Dataset and Results

The dataset used is comprised by a subset from the PICTURE project[1] database, for breast cancer research. In particular, it was used 21 T1-weighted MRI image sets, containing approximately 60 axial slices each, with an average voxel resolution of $0.59 \times 0.59 \times 3$ mm. The total number of artefacts (shoulder copies) present in the dataset is 36. The dataset was annotated by an expert and contains the delineation of each shoulder.

The proposed pipeline (P1234) includes the modules: 1) *Thresholding*, 2) *Region Growing*, 3) Convex Hull and 4) *Minimum Path*. The methods were subsequently introduced to solve specific issues with the base pipeline (P14). To verify the usefulness of each step, experiments concerning the removal of the middle modules were conducted. The sequential improvement can be seen in Figs. 1(d)-1(g) and Table 1.

The evaluation of the results, present in Table 1, was performed using Area: Area Overlap Measure (AOM), the Combined Measure (CM) of AOM, under and over-segmentation [2]; and Contour metrics: Average Minimum Euclidean Distance (AMED), Hausdorff distance (HD) [7].

Table 1: $\mu$ and $\sigma$ results of the segmentation pipelines.

| **Pipelines** | **AOM** | | **CM** | | **AD** | | **AMED** | | **HD** | |
|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 50 | (28) | 66 | (19) | 6.03 | (7.03) | 6.94 | (6.94) | 29.23 | (17.90) |
| P12 | 56 | (29) | 70 | (20) | 4.89 | (6.04) | 5.62 | (6.07) | 25.76 | (17.49) |
| P123 | 62 | (24) | 71 | (21) | 3.11 | (4.59) | 3.82 | (5.09) | 20.59 | (14.43) |
| **P1234** | **64** | **(26)** | **72** | **(24)** | **2.52** | **(3.55)** | **3.15** | **(4.25)** | **18.75** | **(13.57)** |

$P_{xx}$ relate to the module steps, 1) Coarse detection, 2) EntropyRG, 3) Convex Hull, 4) Min. Path.
Metrics units: AOM and CM-% (max is best); AD, AMED and HAUSD-mm (min is best)

The overall results seem to indicate a contour rather close to the Ground Truth (GT), on the best case averaging ($\mu$) 2.52mm (AD) and 3.15mm (AMED). The AOM implies a missing third of the overlapped area that is either of GT, segmentation or crossed mismatch. The growth of this measure correlates with CM. It is noteworthy that CM also slightly increases in the standard deviation ($\sigma$) along the inclusion of modules on the pipeline, which seems to model two trends according to visual inspection. The first is the inclusion of real object's missing portions and the second relates to the inclusion of non-object regions on segmented area. The full pipeline (P1234) perfects the contour in most of the cases but creates some over-segmentation issues, worsening the overall $\sigma$ (AOM and CM), with slightly better $\mu$ and the best contour metrics for all tested pipelines. The concordant larger values of AOM and CM's $\sigma$, and HAUSD ($\mu$ and $\sigma$) suggest that the cases are generally good, with a few cases where the results fail considerably, leaving AD and AMED mostly unscathed.

## 4 Conclusion

The images available presented objects of the highest intensity over all the dataset, despite in several cases having uneven areas near the object's borders. The proposed method comprises modules that progressively narrowed the contour to the correct position. According to the metrics employed, the complete pipeline seems to have been improved over the tested module combinations. However, the dataset's size is limited which does not enable a strong validation. Significant errors rarely occurred and only if the mask was too narrow or far from the desired contour, sufficiently misguiding the *Minimum Path* to over-segment. At times, the algorithm ran over incorrect isolated parts of earlier segmentation. This was mainly caused by a mismatch of the *Minimum Path* positioning due to the biparted processing slice interface. This work tackled an issue that arose from the task of segmenting anatomical structures from MRI data. Despite the AMED of around 3mm (due to over-detection portions), this may be sufficient for accurate results on the pipeline's next step, i.e. detection of other structures. Overall results are very promising, although some path optimization step issues must be revised to remove error propagation.

## 5 Redacted for Review - Acknowledgements

This work is financed by the ERDF - European Regional Development Fund through the Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement and through the Portuguese National Innovation Agency (ANI) as a part of project BCCT.Plan - NORTE-01-0247-FEDER-017688.

## References

[1] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *Int J Comput Vis*, 22(1):61–79, 1997. doi: 10.1023/A:1007979827043.

[2] Matthias Elter, Christian Held, and Thomas Wittenberg. Contour tracing for segmentation of mammographic masses. *Phys Med Biol*, 55(18):5299–5315, aug 2010. doi: 10.1088/0031-9155/55/18/004.

[3] V. Fiaschetti, C.A. Pistolese, V. Funel, M. Rascioni, G. Claroni, F. Della Gatta, E. Cossu, T. Perretta, and G. Simonetti. Breast {MRI} artefacts: Evaluation and solutions in 630 consecutive patients. *Clinical Radiology*, 68(11): e601 – e608, 2013. doi: 10.1016/j.crad.2013.05.103.

[4] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., 2006. ISBN 013168728X.

[5] Hélder P. Oliveira, Jaime S. Cardoso, André T. Magalhães, and Maria J. Cardoso. A 3d low-cost solution for the aesthetic evaluation of breast cancer conservative treatment. *Comput Methods Biomech Biomed Eng Imaging Vis*, 2(2):90–106, dec 2013. doi: 10.1080/21681163.2013.858403.

[6] S.K. Pal and N.R. Pal. Segmentation using contrast and homogeneity measures. *Pattern Recogn. Lett.*, 5(4):293 – 304, 1987. doi: 10.1016/0167-8655(87)90061-4.

[7] Enmin Song, Shengzhou Xu, Xiangyang Xu, Jianye Zeng, Yihua Lan, Shenyi Zhang, and Chih-Cheng Hung. Hybrid segmentation of mass in mammograms using template matching and dynamic programming. *Academic Radiology*, 17(11):1414–1424, nov 2010. doi: 10.1016/j.acra.2010.07.008.

[8] J. F. Teixeira and H. P. Oliveira. Spacial Aliasing Artefact Detection on T1-Weighted MRI Images. In *Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings*, pages 462–470, 2017. doi: 10.1007/978-3-319-58838-4_51.

[9] Andrew Top, Ghassan Hamarneh, and Rafeef Abugharbieh. Active learning for interactive 3d image segmentation. In *Med Image Comput Comput Assist Interv (MICCAI)*, volume 6893 of *LNCS*, pages 603–610. Springer Berlin / Heidelberg, 2011. doi: 10.1007/978-3-642-23626-6_74.

[10] G. W. Zack, W. E. Rogers, and S. A. Latt. Automatic measurement of sister chromatid exchange frequency. *J Histochem Cytochem*, 1977.

[1] http://www.vph-picture.eu/

# Using semi-hidden Markov models and Gaussian mixtures for heart sound segmentation

Jorge Oliveira
Francesco Renna
Miguel Coimbra

Instituto de Telecomunicações, Faculdade de Ciências da Universidade do Porto.

## Abstract

The analysis of heart sounds is a challenging task, due to the quick temporal onset between successive events and the fact that an important fraction of the information carried by phonocardiogram (PCG) signals lies in the inaudible part of the human spectrum. For these reasons, computer-aided analysis of the PCG can improve dramatically the quantity of information recovered from such signals. The first step in the automated analysis of PCG consists in identifying the first heart sound (S1), the systolic period (siSys), the second heart sound (S2), and the diastolic period (siDia). In this work, we use a semi-hidden Markov model, where the emission distribution is modelled using a multi-feature Gaussian mixture model (GMM) and the sojourn times distribution is modelled with Poisson or Gaussian distributions. The corresponding parameters are automatically adapted to each subject through the EM algorithm. On the other hand, multi-dimensional GMMs are used in order to improve the discrimination between feature vectors from different PCG segments. The proposed approaches outperformed the current state-of-the-art solutions when using a subject independent approach. In the Digiscope dataset, and using the Gaussian distribution as an approximation for the sojourn time, our algorithm achieved 89% of positive predictability per state.

## 1 Introduction

The phonocardiogram (PCG) is a cheap and non-invasive tool to assess the mechanical heart functionality [1]. The phonocardiogram contains valuable information, which, if analyzed quantitatively, can lead to important diagnostic results. The first main goal, when processing a PCG signal is to split the PCG into heart cycles, and recognizing its components. Each heart cycle is mainly composed by the first heart sound (S1), the systolic period (siSys), the second heart sound (S2), and the diastolic period (siDia) [1]. PCG signals are usually processed in a specific domain, e.g: the time domain (Shannon energy [2]) or the frequency domain (homomorphic filter [3]), and segmentation is performed using machine learning techniques such as: Artificial Neural Networks (ANN) [4], Support Vector Machines (SVM) [5] and hidden Markov models (HMM) [6]. Due to the sequential nature of heart sound signals, HMMs offer a more natural modeling approach compared to other methods. Recently, HMMs have shown to be very effective in modeling heart sound signals: in Gill et al. [6], the signal is pre-processed and a subset of candidates (peaks) are extracted from the homomorphic envelogram, and these candidates are classified using a discrete-time HMM, where the state distribution is modeled using the time-duration from the preceding candidate to the current one. Chung [7], detected and classified heart sounds using first a left-right HMM model (the first state is assumed to be known) and later a fully connected HMM. The variability in each state is modeled by using multiple mixtures of a Gaussian multivariate distribution. Schmidt et al. [8] implemented a hidden semi Markov model (HSMM) using the homomorphic filtering envelogram as an observation to the system. This has the advantage (compared to the traditional HMM) that every state duration is explicitly modeled in the state transition matrix. The state duration distribution function is modeled by a Gaussian distribution, where the systolic (siSys) and diastolic (siDia) duration parameters are estimated through autocorrelation analysis of the homomorphic filtering envelogram. Springer et al. [9] expanded Schmidt's algorithm mainly on the study of the emission probability distribution. He explored the usage of SVMs to estimate the emission probabilities of the HSMM. In this work we model emission distributions for the different features collected from the PCG signal by using Gaussian Mixture Model (GMM) priors. Such distributions have been proved to be very flexible in representing data from different domains, as they can approximate complex distributions under mild regularity conditions [10], and they have provided state-of-the-art results in various signal processing tasks [11]. Moreover, GMMs allow for simple, and computationally efficient evaluation of the emission probability from data. In this paper, we implemented and improved the Springer et al. [9] algorithm in a subject independent approach. Our proposed algorithm uses E-M routines to adjust the HSMM parameters to each subject. Secondly, the parameters of the emission probabilistic distribution are modelled using a GMM, whose parameters are estimated by using the well-known expectation maximization (EM) algorithm on annotated training data.

## 2 Background

HSMM are an extension of the HMM, namely the sojourn time series are not necessarily geometrically or exponentially distributed over the sojourn time in a state. In other words, the Markovian assumption is relaxed. The future visited states are conditioned not only on the present state, but also for how long it has been in the present state (sojourn time). This allows an arbitrarily sojourn times distribution in any state [12]. The HSMM is specified using the following tuple Θ={Γ,B,π,D}, where B contains the emission probability distribution parameters (for continuous emissions) or it represents the emission probability matrix (for discrete emissions), $\pi$ is the initial state distribution vector, D is the sojourn time distribution matrix and Γ is the state transition probability matrix [12]. We first define $D$ as the sojourn time distribution matrix. The entries of $D$ are $d_{s_k}(u_k)$, the probability of spending $u_k$ units of time in the state $s_k \in \{S1, siSys, S2, siDias\}$. The sojourn time distribution is approached using a Poisson or a Gaussian distribution. Furthermore we define $d^*_{s_k}(u_k)$, as the survivor function of the sojourn time distribution. We also define $r$ as the total number of state transitions that occurred until time n; and finally, we also define as $N(t)$ as the current state at time t. Let us denote by $x_t$ the vector containing samples corresponding to features extracted from the PCG signal at time $t$. Then, modeling emissions via GMMs implies that the multivariate probability density function (pdf) of $x_t$, conditioned on the state $s_k$ is given by:

$$b_{s_k}\left(x_t \middle| \{p_{i,s_k}\}, \{\mu_{i,s_k}\}, \{\Sigma_{i,s_k}\}\right) = \sum_{i=1}^{I} p_{i,s_k} N\left(\mu_{i,s_k}, \Sigma_{i,s_k}\right), \quad (1)$$

where $N\left(\mu_{i,s_k}, \Sigma_{i,s_k}\right)$ represents the multivariate Gaussian pdf with mean $\mu_{i,s_k}$ and covariance matrix $\Sigma_{i,s_k}$, and $I$ is the number of Gaussian components in each, state-dependent, multivariate GMM distribution [13].

## 3 Building and optimizing a model

In our models, we assume that each state of a HSMM corresponds to an element of the heart sound signal because the signal characteristics in each element are thought to be homogeneous. Our models are directed graph and each node has only one parent node with the exception of the first node. For simplicity, this model ignores S3, S4 and murmur sounds. The initial states distribution ( $\pi$ ) are initialized with equal starting probabilities. The Γ matrix is fixed according to the usual state transition $\{S1 \rightarrow siSys \rightarrow S2 \rightarrow siDias\}$. For the purposes of this paper we will ignore any skipped beats, extra sounds or murmurs. The parameters of the emission distribution are inferred from annotated training data, using the EM algorithm for GMM distribution learning [10]. To estimate the initial parameters D, we compute an auto-correlation function over the homomorphic envelogram. From this, we use some heuristics proposed by Schmidt et al. [8] and later by Springer et al. [9, 14], which are heart rate dependent. Our goal, as usual, is to maximize the complete log likelihood, which is a function of the observed variable $X$ (the observation sequence) and the hidden variable $S$ (the state sequence). The likelihood of a state sequence $S$ of a HSMM is:

$$\begin{aligned} P(X,S,\Theta) = \pi &\left\{\prod_{k=2}^{r-1} \gamma_{s_{k-1,k}} d_{s_k}(u_k)\right\} \\ &\times \gamma_{s_{r-1,r}} d^*_{s_r}(u_r) \left\{\prod_{l=1}^{n} b_{s_{N(l)}}(x_l)\right\}, \end{aligned} \quad (2)$$

where $s_k$ is the $k^{th}$ visited state and $u_k$ is the sojourn time of the $k^{th}$visited state. In order to maximize likelihood equation, it is necessary to initialize properly the model parameters. The parameters Θ are estimated using the EM method [15]. The method uses an iterative algorithm that converges to an optimal solution [16].

## 4 Methodology

To train our models, we used the Springer [9] dataset. To test our models, we have used the Digiscope dataset [17]. Following previous literature [2], the system first subtracts the minimum and scales the signal. The scaled signal is filtered using a Butterworth lowpass filter of order 10 with a cutoff frequency of $100Hz$, since the majority of the frequency content of the S1 and S2 (for the DigiScope dataset) is over the range $30-80Hz$. From the filtered signal, several features are extracted: homomorphic envelogram, Hilbert envelogram, wavelet-based features and power spectrum density (PSD) based features. The performance of the HSMM was measured as the model's capacity to recreate the state sequence annotated by the cardiacpulmonologists, namely the positive predictability per state. The testing dataset [17] is composed of samples from 29 different healthy individuals, recorded at 4000Hz in the mitral spot. The training dataset [9] contains 792 healthy and non-healthy adult patients sampled at 1000Hz.

## 5 Results

In our experiments, we have tested two distinct ways to model the sojourn time in the HSMM: the Gaussian and the Poisson distribution. The Gaussian and Poisson distribution do not provide significantly different results, when recreating the "true" state sequence in a heart sound signal (see Figure 3). Our results outperformed significantly the original Springer algorithm in our dataset [9] (81% of $P^{+}_{state}$). This is due to the fact, that Springer algorithm uses general settings, instead of adjusting $\Theta$ to each subject. Using EM routines, our algorithm selects the parameter set that maximizes the complete log likelihood, although it is not guarantee that the global maximum is always achieved. Indeed, the general parameter set estimated by Springer are used as an initialization point. The only parameters that are not adjusted to the subject are the emission distribution parameters, since the corresponding GMM priors are obtained via a supervised approach, i.e., they are inferred from annotated training data.

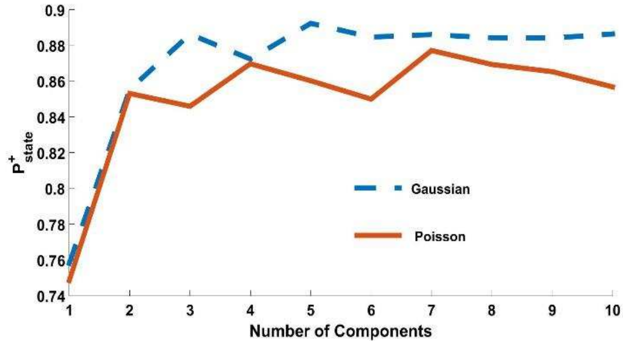


Figure 1: Subject independent $P^{+}_{state}$ results, when using a GMM to model the emission distribution of the HSMMs.

Our $D$ matrix is initialized using heuristics extracted from the auto-correlation function over the homomorphic envelogram. Although the initial $D$ produces good results on average, they performe poorly in some unusual cases, namely in neonate cases, where is common to have systolic times longer than diastolic times. In this situation, our model maps the reverse expected state sequence annotated by the expert. Finally our approaches produced poor results in some infants that were characterized by a hyper-phonetic S2 with respect to S1. In this situations, the S1 segments were not detected by our algorithms.

## 6 Conclusion and Future Work

In this paper, we use GMM priors to model the emission distribution in the HSMM. This model was able to decode the "true" state sequence of events in the PCG signal. Our approach outperformed the Springer algorithm, which is considered the state-of-art in the heart sound segmentation. In our dataset, we achieved 89% of $P^{+}_{state}$ compared to 81% of $P^{+}_{state}$ achieved by the algorithm in [9]. Numerical results seem to hint that the reason for this performance might be related also to the enhanced capability of multi-feature GMM priors in describing signals from the different states and discriminate among them. For future work, we would like to incorporate methods to adapt the GMM distribution online for different subjects.

## Acknowledgment

This work is a result and funded by the project NanoSTIMA (NORTE-01-0145-FEDER-000016), supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF). It is also a result of the internal project SmartHeart in scope of project UID/EEA/50008/2013.

## References

[1] J. E. Hall and A. C. Guyton, Textbook of medical physiology. Philadelphia, Pa.: Saunders/Elsevier, 12th ed., 2011.

[2] H. Liang, S. Lukkarinen, and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," in Computers in Cardiology, pp. 105–108, 1997.

[3] C. N. Gupta, R. Palaniappan, S. Swaminathan, and S. M. Krishnan, "Neural network classification of homomorphic segmented heart sounds," Appl. Soft Comput., vol. 7, no. 1, pp. 286–297, 2007.

[4] T. Leung, P. White, W. Collis, E. Brown, and A. Salmon, "Classification of heart sounds using time-frequency method and artificial neural networks," in Engineering in Medicine and Biology Society, IEEE Conference, vol. 2, pp. 988–991, 2000.

[5] J. Vepa, "Classification of heart murmurs using cepstral features and support vector machines," in Engineering in Medicine and Biology Society, IEEE Conference, pp. 2539–2542, 2009.

[6] D. Gill, N. Gavrieli, and N. Intrator, "Detection and identification of heart sounds using homomorphic envelogram and self-organizing probabilistic model," in Computers in Cardiology, pp. 957–960, 2005.

[7] Y.J. Chung, Pattern Recognition and Image Analysis, Iberian Conference, ch. Classification of Continuous Heart Sound Signals Using the Ergodic Hidden Markov Model, pp. 563–570. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[8] S. Schmidt, E. Toft, C. Holst-Hansen, C. Graff, and J. Struijk, "Segmentation of heart sound recordings from an electronic stethoscope by a duration dependent hidden-Markov model," in Computers in Cardiology, pp. 345–348, 2008.

[9] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression hsmm-based heart sound segmentation.," IEEE Transactions on Biomedical Engineering, vol. 63, no. 4, pp. 822–832, 2016.

[10] H. W. Sorenson and D. L. Aspach, "Recursive Bayesian estimation using Gaussian sums," Automatica, vol. 7, no. 4, pp. 465-479, 1971.

[11] G. Yu, G. Sapiro and S. Mallat, "Solving inverse problems with piecewise linear estimators: from Gaussian mixture models to structured sparsity," IEEE Trans. Image Process., vol. 21, no. 5, pp. 2481-2499, 2012.

[12] S. Zheng Yu, "Hidden semi-Markov models," Artificial Intelligence, 2010.

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

[14] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Support vector machine hidden semi-Markov model-based heart sound segmentation," in Computing in Cardiology Conference (CinC), 2014, pp. 625–628, Sept 2014.

[15] Y. Guédon, "Estimating Hidden Semi-Markov Chains from Discrete Sequences," Journal of Computational and Graphical Statistics, vol. 12, no. 3, pp. 604–639, 2003.

[16] G. D. Forney, "The Viterbi algorithm," Proceedings of the IEEE, vol. 61,no. 3, pp. 268–278, 1973.

[17] D. Pereira, F. Hedayioglu, R. Correia, T. Silva, I. Dutra, F. Almeida, S. Mattos, and M. Coimbra, "DigiScope - Unobtrusive collection and annotating of auscultations in real hospital environments," in Engineering in Medicine and Biology Society, IEEE Conference, pp. 1193–1196, 2011.

# Automatic Segmentation of Breast Complete Contour Using Multi-Modal Data

Hooshiar Zolfagharnasab[1], João P. Monteiro[1], João F. Teixeira[1], Filipa Borlinhas[2], Hélder P. Oliveira [1]
{hooshiar.h.z,jpsm}@ieee.org,{jpfteixeira.eng,filipaborlinhas}@gmail.com, helder.f.oliveira@inesctec.pt

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

[2] Portuguese Oncology Institute of Lisbon (IPOLFG, EPE), Portugal

## Abstract

Providing a planning tool for Breast Cancer Surgery plays an important role to not only increase patient's quality of life after surgery, but also improve surgical techniques to obtain more satisfactory cosmetic outcomes. Such planning tool requires a methodology to segment breast from patient's 3D models automatically. Achieving seamless segmentation demands to detect breast contour flawlessly; however, current methodologies either need user interaction, or detect the contour partially. In this paper, we intend to propose an improved methodology to detect breast contour from patient's complete 3D models. Additionally, a method is proposed to close the segmented breast using additional information obtained form MRI data.

## 1 Introduction

Although breast cancer is known as the second most frequent cancer among females, it can be treated with a 10-year survival rate for 83% of patients [5]. Whereas tumor resection surgery is known as the main invasive treatment, even performing Breast Conservative Surgery (BCS) is led to impose deformations on breast shape, that causes an unpleasant impact on patients' quality of life. Therefore, during the process of planning for the treatment, the existence of planning tool can help medical team to interact more about surgical techniques, while it can also allow patients to participate in the decision process to be informed about the consequences of the treatment. To perform the required action, the aforementioned tool demands a segmented breast; however, current methodologies for breast segmentation require user interaction to select contour endpoints [1], or perform an incomplete detection of breast region [3, 4].

In this paper, a multi-modal approach is proposed to detect complete breast region using Three-Dimensional (3D) data. While the outer border (inferior contour) is segmented from 3D reconstructed data, the interior one (pectoral muscle) is imported from MRI data.

## 2 A Brief History on Breast Region Segmentation

State-of-the-art methodologies mostly used 2D images to find the inferior contour, which connects the two endpoints, from the sternum in one side to the vicinity of axilla in the other side. Using frontal RGB images, Cardoso *et al.* [1] found the lower half contour by computing the shortest path between the mentioned endpoints. Their method was based on a semi-automatic approach, since some key points were manually identified: breast contour endpoints and nipples position. Subsequently, they improved their approach for an automatic and correct detection of the contour. They exploited the circular shape of breast to solve the shortest path problem in polar coordinate. Nonetheless, the detected contour was too sensitive to the gradient changes of the used RGB images [3, 6].

Considering catenary curve as shape constraint, active contours were adopted in the research performed by Lee *et al.* [3]. The inferior contour was detected within two stages: coarsely fitting an optimized catenary curve to the orientation of breast, and then applying a proper balloon force to collapse the coarsely fitted curve to the breast contour. Whereas their initial model was an open curve, the above half of frontal contour (upper contour) was ignored.

Aiming to get the benefits from 3D local information, Oliveira *et al.* [4] proposed an approach using depth data acquired with a Microsoft Kinect device. To find the inferior contour, they employed a template matching to find the fiducial points followed by a shortest path approach in polar coordinates (centered at breast *breast peak*). Note that the required breast *breast peak* which corresponds to the point in the breast where disparity attains the lowest value, can be found by filtering gradient vector field via a template. The discontinuities in the upper part was defined by the horizontal line connecting the two contour sides.

## 3 Multi-modal Breast Segmentation

In this paper, it is proposed to obtained the complete breast contour using a multi-modal scenario. In one side, the exterior contour is detected from reconstructed 3D surface information obtained with the algorithm of Costa *et al.* [2]. On the other side, the interior contour (which is pectoral muscle) is obtained from MRI data via annotating them by experts manually. Putting the work of Oliveira *et al.* [4] as the baseline, the proposed algorithm can be implemented as a sequence of a few high-level operations, including detection of exterior contour (inferior and upper) from RGB-D data, and interior contour (pectoral muscle) from MRI data.

### 3.1 Exterior Breast Contour Segmentation

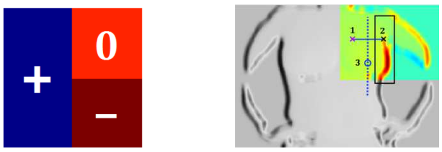
Contrarily to baseline method [4], the proposed methodology is designed to use 3D data as input. Hence, initially, 3D data are projected into the coronal plane to provide depth information.

The detection of inferior breast contour has been inspired from the baseline method, where the breast contour was detected by finding the shortest path which connects neighboring pixels of a cost map. Therefore, a cost map is generated via polar conversion of breast tear-drop shape, and manipulated with an appropriate weight function to highlight the contour pixels. As in [4], similar technique was employed to detect the breast peak point. However, a new approach (based on template matching) is proposed to detect the two endpoints of the inferior contour to overcome the drawback of the baseline method. Detection of the first endpoints (external endpoint) is carried out by finding it near to the axilla, through convolving the designed kernel (depicted in Figure. 1(a)) and filtering the topmost extremity of its response. Followed by the external endpoint, the internal one is located at the same distance but in the opposite side of the breast vertical median plane crossing through the *breast peak* point. Figure. 1(b), depicts the triple detected aforementioned points for the left breast. Flipping the kernel horizontally makes it possible to detect the corresponding points, but on the right breast.

The two endpoints are also assists the detection of breast upper contour; however, since the region where upper breast contour exists, is characterized as an edge-less region, using an approach employed to outline the lower contour would lead to a wrong detection. To solve this, the values of the cost map are rectified by element-wise multiplication in $\alpha_{r,c}$.

$$\alpha_{r,c} = \begin{cases} 1 & r-w \le a \times c + b \le r+w \\ 0 & otherwise \end{cases} , \qquad (1)$$

where $r,c$ are rows and columns of cost map (in polar coordinates) respectively, and $a \times c + b$ is the line which connects the endpoints. This



(a) (b)

Figure 1: a) Suggested template for detection of left external endpoint, b) response of the template matching on the left breast RoI is shown colorful. The external endpoint (labeled 1, or $s_{beginning}$) is used together with breast peak point (labeled 3) to locate the internal endpoint (labeled 2, or $s_{ending}$).

multiplication imposes a tunnel with initial width of $w = 1$ pixel between the endpoints where the upper contour lies upon. Increasing the width of tunnel iteratively assures the detection of the shortest path which connects the mentioned endpoints through the tunnel.

### 3.2 Interior Breast Contour Segmentation

Providing only surface, RGB-D data are insufficient to assist the detection of interior breast contour (pectoral muscle). Therefore the missing contour is provided by MRI data instead. However, since RGB-D and MRI data are captured in different modalities, a transformation must be performed to bring one's coordinate to another. The transformation of MRI data into RGB-D may be criticized due to the compression of breast tissue during MRI acquisition. Nevertheless, problem simplification leads to assume that unlike the breast tissue, the pectoral muscle is reluctant to deform during the acquisition. Also, it is adjudged that changing in patient's position (from prone to upright) has no effect in the rigidity of pectoral muscle. Having assumed the simplifications, the pectoral muscle tends to keep its relative position with respect to the skin (sternum region). These assumptions provide a strategy to define a framework for the required transformation. In this way, two solutions are proposed, in which require the definition of 4 fiducial points, in both data, which are assumed to be in the same locations on the patient's torso.

#### 3.2.1 Fitting planes to the corresponding fiducial points

In this solution, 4 planes are fitted to the regions defined by the 4 fiducial points in the RGB-D space. These planes play a guide role for the correct positioning of the pectoral muscle during an iterative process.

#### 3.2.2 Semi-rigid transformation of MRI data

The second solution is inspired by Iterative Closing Points (ICP) algorithm in which transforms the pectoral muscle (MRI data) to RGB-D space in an iteratively procedure by coarsely approaching the fiducial points. During the iterations, the required parameters for rotation, translation and scaling are determined, which are consecutively applied to MRI data until the average distance between transformed models becomes less than 0.1 millimeters.

## 4 Results and Discussion

The exterior and the interior contours are evaluated quantitative and qualitative, respectively. A database consisting of 32 patients was used for the evaluation. Each patient 3D reconstructed model was obtained using the framework developed by Costa *et al.* [2]. Besides, pectoral muscle information was obtained from MRI slices which were manually annotated by an expert.

### 4.1 Exterior Breast Contour Segmentation

Evaluation is performed using the Hausdorff and average distance between the ground-truth contour and the one obtained from the proposed methodology (see Table 1). The Hausdorff distance between point sets A and B is defined as:

$$h(A,B) = \underset{a\in A}{max}\ \underset{b\in B}{min}\ \parallel a-b \parallel \ , \qquad (2)$$

Table 1: Breast contour detection error (in mm).

| | Detected → Ground-truth | | Ground-truth → Detected | |
|---|---|---|---|---|
| | Average | Hausdorff | Average | Hausdorff |
| **Mean** | 3.84 | 11.51 | 2.64 | 9.00 |
| **Std** | 1.41 | 4.77 | 0.95 | 3.92 |
| **Max** | 7.23 | 23.08 | 5.54 | 20.18 |
| **Min** | 1.69 | 5.45 | 1.02 | 3.84 |

where $\parallel . \parallel$ is the Euclidean distance representing the worst case scenario. The average error of 3.84*mm* from detected contour to the ground-truth indicates that the exterior contour has been found in a reasonable region.

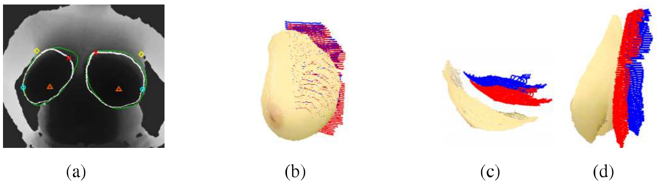


Figure 2: a) Detected exterior contour (white) ,and ground-truth (green), breast peaks (triangle), infra-mammary endpoint (blue circle), internal endpoints (red cross), and external endpoints (yellow diamond) b,c) perspective and top view, and d) lateral view of pectoral muscle transforming suggestions; the suggestion powered by fitting planes is shown in blue while the output of semi-rigid transformation solution is shown in red.

As long as the visual evaluation of inferior contour indicates a competitive detection comparing with the ground-truth, there is a small visible gap between the discovered upper contour and the manually annotated one (see Figure. 2(a)). Despite affecting numerical evaluation, the detected contour is still found in the region that is consensually accepted as upper contour.

### 4.2 Interior Breast Contour Segmentation

The lack of ground-truth for the interior contour, makes it possible to perform only a qualitative evaluation, through the visualization of pectoral muscle position against the location of the breast. It has been shown that the pectoral muscle is located in an accepted anatomical location toward both the breast tissue and sternum bone, mainly using the second approach (Semi-rigid transformation of MRI data). Figure. 2(b), Fig. 2(c) and Figure. 2(d) depict the visual output for one of the patients of the database.

## 5 Conclusion

In this paper, a multi-modal approach is proposed to segment complete contour of breast automatically. The contributions include suggesting a improved methodology to detect inferior contour, and correct delimitation of upper, and interior contours (pectoral muscle). Quantitative and visual inspections of the results demonstrates a good performance and robustness for a wide variety of patients. Future work will focus on the use of the outcomes incorporated in a BCS planning tool.

## Acknowledgment

This work is financed by the ERDF – European Regional Development Fund through the Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement and through the Portuguese National Innovation Agency (ANI) as a part of project BCCT.Plan - NORTE-01-0247-FEDER-017688, and also by Fundação para a Ciência e a Tecnologia (FCT) within PhD grant number SFRH/BD/97698/2013.

## References

[1] J. S. Cardoso and M. J. Cardoso. Breast contour detection for the aesthetic evaluation of breast cancer conservative treatment. *Computer Recognition Systems*, 2(1):518–525, 2007.

[2] P. Costa, J. P. Monteiro, H. Zolfagharnasab, and H. P. Oliveira. Tessellation-based coarse registration method for 3d reconstruction of the female torso. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 301–306. IEEE, 2014.

[3] J. Lee, G. S. Muralidhar, G. P. Reece, and M. K. Markey. A shape constrained parametric active contour model for breast contour detection. In *34th International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4450–4453, 2012.

[4] H. P. Oliveira, J. S. Cardoso, A. T. Magalhães, and M. J. Cardoso. A 3d low-cost solution for the aesthetic evaluation of breast cancer conservative treatment. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, 2(2):90–106, 2014.

[5] American Cancer Society. American cancer society: Breast cancer facts and figures 2015-2016. *American Cancer Society (ACS).*, 2015.

[6] L. Zhao, A. Cheong, G. P. Reece, M. C. Fingeret, S. K. Shah, and F. A. Merchant. Inferior breast-chest contour detection in 3-d images of the female torso. *IEEE Journal of Translational Engineering in Health and Medicine*, 4 (1):1–10, 2016.

# Towards an Automated Detection and Quantification of Mice Arthritis

João Santos[1]
up201605317@g.uporto.pt
Kelwin Fernandes[2]
kafc@inesctec.pt
Jaime S. Cardoso[2]
jsc@inesctec.pt

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

## Abstract

Arthritis is a degenerative disease that affects a large segment of the population with negative consequences on their daily life including their working capabilities. Current medical research on the effect of certain drugs for arthritis is done with mice. In this work, we addressed the problem of segmenting the mice's paws on bottom-view grayscale images for the automatic quantification of arthritis.

## 1 Introduction

Arthritis is a degenerative disease that affects a large segment of the population, being two-thirds of the affected population of working age (18-64 years). Its symptomps include swelling, pain, stiffness and decreased range of motion. As a consequence, the daily activity of these patients gets severely affected.

In order to understand the impact of certain drugs on the treatment against arthritis, research have been conducted on the motion analysis of mice under the effect of such treatments [2, 3]. The procedure starts by inducing arthritis to the mouse on one of its rear legs using medical procedures. Then, the developed drug is administered to the animal under observation in order to quantify changes on the motion patterns of the animal (i.e. asymmetry in the pressure applied by the rear paws). With the purpose of creating a standard evaluation setting, the medical inverstigators built an acrylic crate illuminated from one side. A video camera was positioned under the cages in order to capture the mouse paws primarily. Figure 1 shows some input images. After recording the mice for a period of several minutes, the experts analyse the results by selecting frames where the paw is maximally pressed against to the surface. Then, the paws in the frame are manually segmented and the paw surface's area in touch with the cage is computed. This process is both slow and inaccurate as many frames should be processed in each video in order to compute statistically significant results. Also, the conclusions might be difficult to replicate since the selection of the frames and segmentation of the paws is user dependent. In order to optimize this process, computer vision techniques were designed in this work in order to automate this process. The main goal of this project is to provide objective metrics about the mice behavior (e.g. motion quantification, speed, gait asymmetry, etc.). As a preliminary step, we designed a system to automatically recognize the mouse in the video and to segment the paws.

Several machine learning and computer vision methodologies were proposed in the past for the automatic understanding of mice behavior [1, 4, 6, 8]. However, most of these attempts considered a top-view of the mice and used depth-cameras [6, 8]. Since we're interested in capturing information that can only be seen below the mice, through a transparent surface, using depth information isn't possible in this project. Other strategies that don't rely on depth information used on grayscale and color images, focusing on the extraction of high-level features for the later recognition of actions using classification models [1, 4].

## 2 Methodology

In this work, we focus on the aforementioned acquisition setting (see Figure 1). The acquired videos have different durations, depending on the mice's motion. Also, there are several changes and artifacts in the video that turn the automation process difficult. For instance, the position of the walls in the cage may change from video to video (see Figure 1). Thereby, the first step in the proposed methodology is to learn and to remove the background elements in each video. Since the illumination conditions in the video were strongly controlled, we used a static and very simple background model. In this sense, we apply a Gaussian blur on each frame in order to diminish the impact of noisy pixels. Then, we calculate the mean of every frame in the video. Equation (1) defines the background model, where $\mathcal{V}$ is the input video, understood as a sequence of frames $f \in \mathcal{V}$.

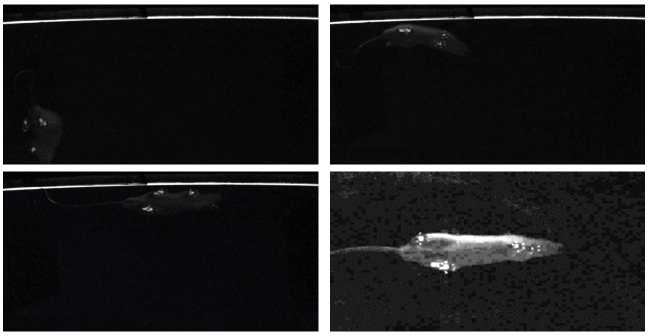


Figure 1: Images from the input videos after a white balance filter for visualization purposes

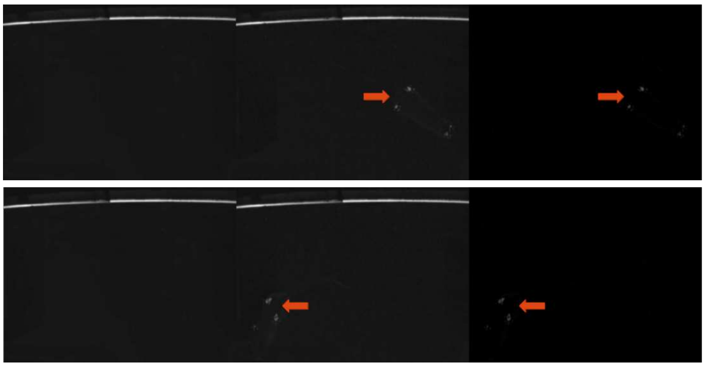


Figure 2: **Left:** background models. **Middle:** Input frame. **Right:** Foreground scores. Paws are point using red arrows for visualization purposes.

$$BG(\mathcal{V}) = \frac{1}{|\mathcal{V}|} \sum_{f \in \mathcal{V}} \mathcal{G}(f) \tag{1}$$

Given that the recognition process is done offline, a good estimation of the background can be precomputing using the entire video, reducing the problem of unstable background models at the very first frames of the video. Then, in order to remove the background regions from the video, the foreground is computed using the absolute of the arithmetic difference between the calculated background image and the frames in the video (see Equation (2)). Figure 2 shows the background model and the corresponding foreground image for several videos.

$$FG(\mathcal{V}, f) = |f - BG(\mathcal{V})| \tag{2}$$

Then, in order to detect the mice, we consider the top $K\%$ of the pixels with highest probability of being foreground (i.e. largest absolute difference) and, after removing small blobs using morphological operators [7], the largest connected component is assigned as a mouse candidate. A connected component is formed by a group of adjacent pixels with the same intensity so the biggest one is the mouse. We considered several alternative approaches for the detection of the mouse (see Figure 3), ranging from OTSU thresholding on the foreground score [9] to clustering

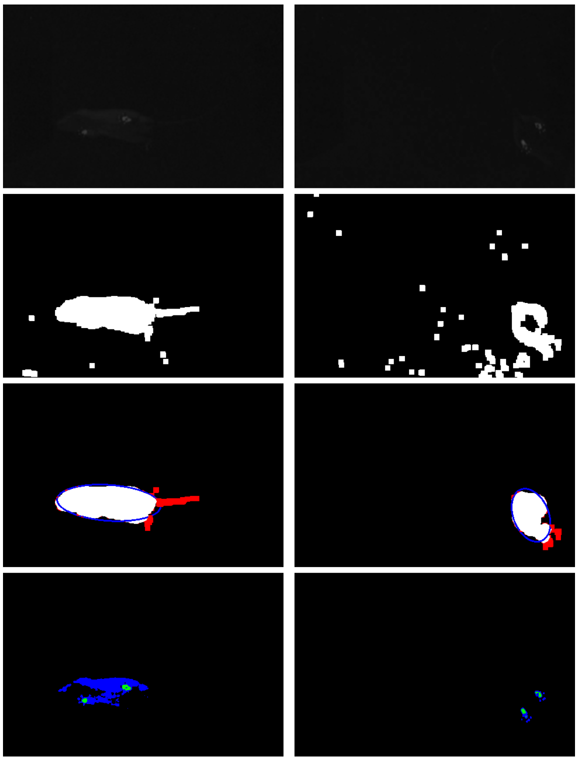


Figure 3: **First row:** input images. **Second row:** foreground. **Third row:** Ellipse (blue) and tail removal (red). **Fourth row:** paws segmentation.

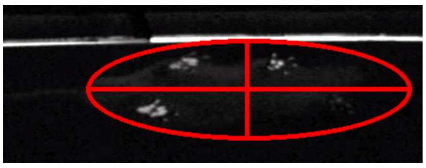


Figure 4: Recognized ellipse with the segmentation of the four main quadrants.

techniques such as K-means [5]. However, the aforementioned approach achieved competitive results with low computational resources.

Since mice can be approximated by an ellipse when they are walking, we fit an ellipse to the largest connected component so we can remove the further parts like the tail or even the head. As a result, we can limit the space search for the paws considerably (see Figure 3). During the videos, the researchers manipulate the mice in order to promote certain displacements and to keep it in the field of view of the camera. This introduced an undersirabled effect in the videos consisting on an abrupt change of the foreground. In order to remove these frames, we compute the expected location of the mouse given the average displacement vectors from the last frames, where a learning decay factor is used to give higher relevance to recent motion patterns. If the expected location of the ellipse is far from the detected location, the frame is discarded. Since these perturbations are not frequent and their duration is very limited, we can remove these frames without affecting the overall performance of the framework through the video.

Since we are only interested in the rear paws due to the medical protocol, we compare every frame with his last and divide the ellipse in front and back, the back being the closest to the previous frame (see Figure 4). Finally, the paws are segmented using $K$-means on the pixel intensity space [5]. $K$-means is a clustering algorithm that aims to partition the observations into $K$ clusters, where each cluster is represented by the average location of its nearest pixels. We used $K = 3$ such that one cluster should cover the background pixels, a second cluster represents the body of the mouse and the third cluster is mainly formed by the mouse paws. Figure 3 shows the segmentation results.

## 3 Conclusions

In this work we addressed the problem of segmenting the mice's paws on bottom-view grayscale images for the automatic quantification of arthritis. Current studies done with this type of images is very time consuming and subjective since it is done manually.

Since this work is on its preliminary stages, subjective evaluation has been conducted on several videos. In the near future, we expect to manually label the videos in order to measure in an objective manner each stage of the aforementioned pipeline. So far, the visual inspection of the outcomes show promising results. Even considering the wrongly recognized frames, the results are very accurate, allowing a fast selection of the representative frame in a more objective manner than by manual sampling and segmentation.

Also, since we are interested on the motion patterns of the mice when walking, we plan to include an action recognition module in order to discard frames where the mouse is in static activities such as sitted or standing on the rear legs. With the current strategy, these situations arise misleading results given that the mice are poorly approximated by the ellipse.

## Acknowledgements

This work was funded by the Project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF), and also by Fundação para a Ciência e a Tecnologia (FCT) within PhD grant number SFRH/BD/93012/2013.

## References

[1] Carlos Fernando Crispim-Junior, Fernando Mendes de Azevedo, and José Marino-Neto. What is my rat doing? behavior understanding of laboratory animals. *Pattern Recognition Letters*, 2017.

[2] Joana Ferreira-Gomes, Sara Adães, and José M Castro-Lopes. Assessment of movement-evoked pain in osteoarthritis by the knee-bend and catwalk tests: a clinically relevant study. *The Journal of Pain*, 9 (10):945–954, 2008.

[3] Joana Ferreira-Gomes, Sara Adaes, Jana Sarkander, and José M Castro-Lopes. Phenotypic alterations of neurons that innervate osteoarthritic joints in rats. *Arthritis & Rheumatology*, 62(12):3677–3685, 2010.

[4] Hueihan Jhuang, Thomas Serre, Lior Wolf, and Tomaso Poggio. A biologically inspired system for action recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. Ieee, 2007.

[5] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[6] João P Monteiro, Hélder P Oliveira, Paulo Aguiar, and Jaime S Cardoso. A depth-map approach for automatic mice behavior recognition. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 2261–2265. IEEE, 2014.

[7] Philippe Salembier, Albert Oliveras, and Luis Garrido. Antiextensive connected operators for image and sequence processing. *IEEE Transactions on Image Processing*, 7(4):555–570, 1998.

[8] Zheyuan Wang, S Abdollah Mirbozorgi, and Maysam Ghovanloo. Towards a kinect-based behavior recognition and analysis system for small animals. In *Biomedical Circuits and Systems Conference (BioCAS), 2015 IEEE*, pages 1–4. IEEE, 2015.

[9] Xiangyang Xu, Shengzhou Xu, Lianghai Jin, and Enmin Song. Characteristic analysis of otsu threshold and its applications. *Pattern recognition letters*, 32(7):956–961, 2011.

# A method for the segmentation of the anterior fascia of the abdominal muscle

Ricardo J. Araújo
ricardo.j.araujo@inesctec.pt
Hélder P. Oliveira
http://www.inescporto.pt/~hfpo/

INESC TEC
Campus da FEUP, Rua Dr. Roberto Frias, 4200-465,
Porto, Portugal

## Abstract

The segmentation of the anterior fascia of the abdominal muscle is an important step towards the analysis of abdominal vasculature. It may advance Computer Aided Detection tools that support the activity of physicians who study vessels for breast reconstruction using the Deep Inferior Epigastric Perforator flap. In this paper, we propose a two-fold methodology to detect the anterior fascia in Computerized Tomographic Angiography volumes. First, a slice-wise thresholding is applied and followed by a post-processing phase. Finally, an interpolation framework is used to obtain a final smooth fascia detection. We evaluated our method in 20 different volumes, by calculating the mean Euclidean distance to manual annotations, achieving subvoxel error.

## 1 Introduction

In the United States, breast cancer is the leading cause of cancer death in women aged 20 to 59 years, being only surpassed by lung cancer in higher ages, a tendency that is observed worldwide [6]. The mastectomy, a surgical procedure where the whole breast is removed, is still frequently performed and has even been increasing in some institutions [3]. Fortunately, the psychological burden of such procedure may be alleviated with the reconstruction of the breast. Among the available options, the Deep Inferior Epigastric Perforator (DIEP) flap has become the state-of-art technique for autologous tissue based breast reconstruction [2]. In this procedure, skin, fat and vessels are moved from the abdominal region to the chest, without weakening the rectus abdominis muscle, also known as abdominal muscle. The harvested vessels are the DIEPs, regularly known as perforators. They have origin in bifurcations of the Deep Inferior Epigastric Arteries (DIEAs) and then perforate the abdominal muscle, heading to the superficial tissues of the abdomen (see Figure 1 for a representation of the local anatomy).

As microsurgery techniques are involved in this type of harvest, medical imaging has been used for preoperative planning. A physician describes the existing perforators, since the viability of the flap is related to the features of the extracted perforator(s) [5]. Characteristics from both the subcutaneous and intramuscular portions of the perforators are taken into account. The anterior fascia of the abdominal muscle separates both of these regions, hence its automatic segmentation facilitates the use of computer based routines to automatically retrieve the required measures. Furthermore, the automatic detection of the fascia would help to determine the origin of each perforator subcutaneous course, which is required to create an accurate map of the dissection locations.

Organs, such as the liver and pancreas, have been segmented using manually labeled atlases and the graph cuts method [7]. However, the abdominal muscle has not been targeted by other authors. In this paper, we describe a semi-automatic method for the segmentation of the anterior fascia of the abdominal muscle, in the region where the perforators arise, based on [1].

## 2 Abdominal muscle anterior fascia segmentation

The relevant region for DIEP analysis is given by the volume that includes the end of each perforator and the locations where the DIEAs enter the posterior lamella of the abdominal muscle (see Figure 1). A margin was considered to avoid losing segments of tortuous vessels.

In terms of image intensities, the fascia cannot be distinguished from the abdominal muscle. Hence, it is considered to be the boundary between this muscle and the subcutaneous region, characterized by a transition from pixels with low intensity (subcutaneous region) to pixels with higher intensity (muscle), which exists over all the columns of each axial slice of the volume of interest (see Figure 2).

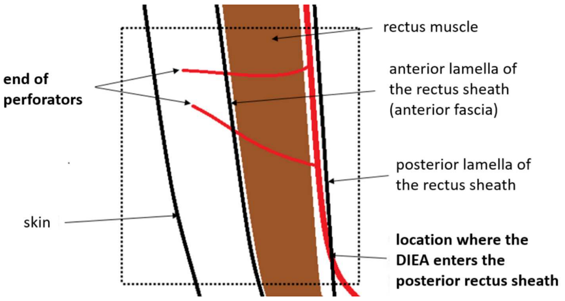

rectus muscle
end of perforators
anterior lamella of the rectus sheath (anterior fascia)
posterior lamella of the rectus sheath
skin
location where the DIEA enters the posterior rectus sheath

Figure 1: Sagittal view representation of the abdominal wall anatomy. The box delimits the region of interest.

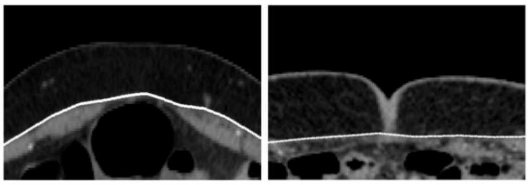


Figure 2: Example axial slices of the volume of interest. Manual annotations of the fascia are shown in white.

### 2.1 Preliminary segmentation

To obtain a preliminary segmentation of the anterior fascia, we conducted some processing steps. Let $B_s$ be the binary image after step $s$. Figure 3 shows the binary images after step $s$ with respect to the examples present in Figure 2. The involved steps are described below:

1. **Otsu's thresholding:** The Otsu's method [4] was used to obtain a threshold that distinguishes the muscle from the subcutaneous region. Occasionally, there are structures that also respond to the threshold and appear connected to the anterior region of the muscle, such as perforators and umbilical tissue. The last modules of the pipeline address these unwanted detections.

2. **Skin removal:** The regions of the original image, where the intensities are equal to zero, are extracted (they include the region outside the body of the patient and cavities filled with air). Among those regions, the area outside the body of the patient is obtained by keeping the largest connected component. Its dilated version is used to remove the skin object present in $B_1$.

3. **Largest connected component:** The largest connected component that exists over all the columns of $B_2$ is selected. If there is none, we iteratively decrease the threshold until such requirement is satisfied.

4. **Filling operation:** Regions lying below the biggest connected component are filled.

5. **Umbilical tissue removal:** Some images have umbilical tissue connecting the skin and muscle regions. When that occurs, it is common that $B_4$ includes an unwanted detection of part of the umbilical tissue (see the bottom row of Fig. 3). To detect slices where this happens, we analyze whether the skin object removed in step 2 is adjacent to $B_4$. This is true if the logical OR operation between them creates a single object. If this is the case, the horizontal derivatives of the OR image are obtained through the Sobel

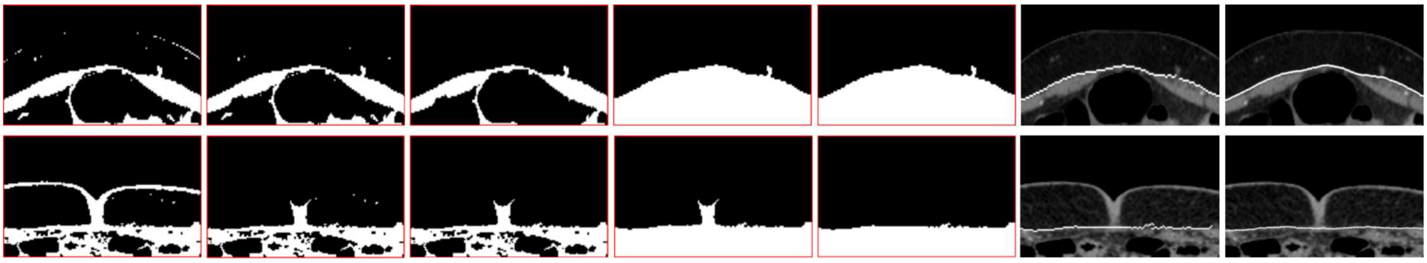


Figure 3: Fascia detection in the two example slices provided in Fig. 2. From left to right, $B_1, \ldots, B_6$, and the final fascia segmentation.

operator, which produces high responses at the isthmus that links the skin and $B_4$. From those detections, a rectangular mask is created and used to remove the connection, producing $B_5$. If not, $B_5$ is equal to $B_4$.

6. **Preliminary fascia detection:** Vertical transitions are obtained for each column of $B_5$. Connected contours, considering 8 neighbors, smaller than $n$ pixels, where $n$ was empirically set to 11, are discarded. We lose the influence of vertical oriented structures which might be still connected to the segmentations, such as vessels.

### 2.2 Final segmentation

To obtain a complete and smooth result, we use a regression framework that takes as input the preliminary fascia segmentation. In sagittal slices, the boundary between the muscle and the subcutaneous region is usually very smooth. For each row of each sagittal slice of our volume of interest, a new fascia point $(p_{row}, p_{col})$ is estimated as:

$$p_{col} = P(p_{row}) \tag{1}$$

where $P$ is a local regression model based on Tukey's bisquare objective function taking into account the preliminary fascia detections contained in the range $[p_{row} - n, p_{row} + n]$, being $n$ expressed by:

$$n = k \cdot \frac{m}{s} \tag{2}$$

where $s$ is the distance in mm between consecutive pixels, characteristic of the volume (same in every direction of the volume after interpolation of data), $m$ is the size of the biggest structures to be neglected, also in mm (vessel with largest caliber in the dataset) and $k$ is a constant. This last parameter can be seen as the amount of data which has to be considered to remove the influence of a certain structure. In this work, $m = 5$ was considered, and $k = 5$ was empirically obtained. We used this interpolation method because it is less influenced by outliers. The rightmost column of Figure 3 shows the final segmentations produced by our methodology.

## 3 Results

The Breast Unit of Fundação Champalimaud provided CTA volumes from 20 different patients. For each volume, the end of each perforators and the locations where the DIEAs perforate the posterior lamella of the abdominal muscle were provided, such that the volumes of interest could be determined. For each axial slice of each volume of interest, a manual annotation of the anterior fascia was performed by an expert. Following the proposed methodology, we obtained 20 fascia segmentations and measured the Euclidean and Hausdorff distances to the corresponding manual annotations. Table 1 shows the mean, best and worst performances achieved.

The voxel spacing varies between different volumes, from 0.7 to 0.9 mm. Thus, the proposed method was able to provide segmentations whose mean Euclidean distance to the manual annotations was lower than the spacing between consecutive voxels. The relatively low mean Hausdorff distance shows that the detections were stable. There was a single case (worst in Table 1) where the resulting segmentation in a particular region of the volume was erroneous. This occurred because in that region, the preliminary fascia segmentation contained a high number of consecutive misdetections, and the robust regression was not able to provide an accurate result. The experiments were run in an Intel Core i7-4500U CPU 1.80@2.40GHz using MATLAB R2014a, and took, in average, 636 s for each volume.

Table 1: Evaluation of the proposed methodology. The Euclidean (E) and Hausdorff (H) distances were calculated between each volume final fascia segmentation and the corresponding manual annotation. Mean, best and worst performances are shown.

| case | E distance (mm) | | H distance (mm) | |
|---|---|---|---|---|
| | GT → seg | seg → GT | GT → seg | seg → GT |
| mean | $0.49 \pm 0.33$ | $0.51 \pm 0.40$ | $1.52 \pm 0.76$ | $1.63 \pm 1.15$ |
| best | 0.28 | 0.28 | 0.79 | 0.79 |
| worst | 1.78 | 2.15 | 4.25 | 6.13 |

## 4 Conclusion

In this paper, we described a method to segment the anterior fascia of the abdominal muscle, which is relevant for the analysis of DIEP. Intensity thresholding and post-processing are used to obtain a preliminary fascia segmentation. Then, a robust interpolation framework is conducted to produce a smooth detection of the fascia, without interferences from neighbor structures. Our method achieved promising results since it produced segmentations whose mean Euclidean distance to the manual annotations was lower than the distance between consecutive voxels.

## Acknowledgments

This work is financed by the ERDF - European Regional Development Fund through the Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement and through the Portuguese National Innovation Agency (ANI) as a part of project BCCT.Plan - NORTE-01-0247-FEDER-017688, and also by Fundação para a Ciência e a Tecnologia (FCT) within PhD grant number SFRH/BD/126224/2016.

## References

[1] R.J. Araújo and H.P. Oliveira. Segmentation of the rectus abdominis muscle anterior fascia for the analysis of deep inferior epigastric perforators. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 537–545, 2017.

[2] A. Cina, M. Salgarello, L. Barone-Adesi, P. Rinaldi, and L. Bonomo. Planning breast reconstruction with deep inferior epigastric artery perforating vessels: multidetector ct angiography versus color doppler us. *Radiology*, 255(3):979–987, 2010.

[3] A.E. Dragun, B. Huang, T.C. Tucker, and W.J. Spanos. Increasing mastectomy rates among all age groups for early stage breast cancer: a 10-year study of surgical choice. *Breast J.*, 18(4):318–325, 2012.

[4] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. Syst., Man, Cybern., Syst.*, 9(1):62–66, 1979.

[5] T.J. Phillips, D.L. Stella, W.M. Rozen, M. Ashton, and G.I. Taylor. Abdominal wall ct angiography: a detailed account of a newly established preoperative imaging technique. *Radiology*, 249(1):32–44, 2008.

[6] R.L. Siegel, K.D. Miller, and A. Jemal. Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1):5–29, 2015.

[7] T. Tong, R. Wolz, Z. Wang, Q. Gao, K. Misawa, M. Fujiwara, K. Mori, J.V. Hajnal, and D. Rueckert. Discriminative dictionary learning for abdominal multi-organ segmentation. *Med. Image Anal.*, 23(1):92–104, 2015.

# The Influence of Image Normalization in Mammographic Classification with CNNs

Ana C. Perre
ana.perre@ipcb.pt
Faculdade Ciências da Saúde,
Universidade da Beira Interior,
Covilhã, Portugal

Luís A. Alexandre
lfbaa@ubi.pt
Dept. Informática,
Universidade da Beira Interior,
Covilhã, Portugal

Luís C. Freire
luis.freire@estesl.ipl.pt
Escola Superior de Tecnologia da Saúde de Lisboa,
Instituto Politécnico de Lisboa
Lisboa, Portugal

## Abstract

In order to improve the performance of Convolutional Neural Networks (CNN) in the classification of mammographic images, many researchers choose to apply a normalization method during the pre-processing stage. In this work, we aim to assess the impact of six different normalization methods in the classification performance of two CNNs.

Results allow us to concluded that the effect of image normalization in the performance of the CNNs depends of which network is chosen to make the lesion classification; besides, the normalization method that seems to have the most positive impact is the one that subtracts the image mean and divide it by the corresponding standard deviation (best AUC mean with CNN-F = 0.786 and with Caffe = 0.790; best run AUC result was 0.793 with CNN-F and 0.791 with Caffe).

## 1 Introduction

Mammographic images are interpreted by highly trained radiologists. However, due to the frequent need of analyzing large amounts of images which are produced daily in medical institutions, they may misinterpret between normal and abnormal tissues [1]. Therefore, it is important to develop automatic or semi-automatic computer-assisted tools that can help radiologists in the detection and interpretation of suspicious regions on mammograms [2]. Convolutional Neural Networks (CNN) have been recently successfully used in the medical field for detection and classification of mammographic lesions [1, 2, 3].

To improve the performance of CNNs in this task , many researchers choose to apply a normalization pre-processing method to mammographic images in the pre-processing stage [1], which is justified by the fact that images are obtained with different exposure conditions and are affected by noise and some artifacts[2]. Furthermore, to perform an accurate analysis, it is necessary to achieve an optimal image contrast [2].

In the paper of [4], the authors found that the use, or not, of a preprocessing image normalization method could yield to different performances of the classification tool. Therefore, in this paper we intend to deepen understand such impact by using six different image normalization methods, being the first four methods variations of Global Contrast Normalization (GCN). Therefore, we have: (method 1) subtracting the image mean; (method 2) subtracting the image mean and dividing by the standard deviation; (method 3) Histogram equalization; (method 4) Histogram equalization in combination with method 2. (method 5) and (method 6) used the same GCN applied in method 1 and method 2, respectively, in combination with a local contrast normalization (LCN). Lastly, we tested the classification process on the same images without normalization, which we call "NoNORM" (see fig. 1 and 2).

## 2 Related Work

Rouhi *et al.* [2] used local area histogram equalization, that stretched the intensity of image pixels to extend the contrast, and then the median filtering, that is a nonlinear operation used to reduce noise ("salt and pepper" and speckle noise). The mammographic images has been normalized and whitened in Jiao *et al.* [1], in the first step the dataset was normalized to the range [0,1] by subtracting them by their mean, and in the second step, they used a method named PCA whitening by dividing the standard deviation of its elements. Arevalo *et al.* [3] applied two normalization types: (1) Global Contrast Normalization, by subtracting the mean of the intensities in the image to each pixel (the mean is calculated per image, not per pixel), and (2) Local Contrast Normalization, that mimics the behavior of the visual cortex and reduces statistical dependencies, which accentuates differences between input features and accelerates gradient-based learning.

## 3 Material and Methods

The dataset used was the BCDR-FM dataset (Film Mammography Dataset) from Breast Cancer Digital Repository [1]. The downloaded subset, named BCDR-F03 - "Film Mammography Dataset Number 3", comprises 736 grey-level digitized mammograms (426 benign and 310 malign mass lesions) from 344 patients. These are distributed into Medio-Lateral Oblique (MLO) and Cranio-Caudal (CC) views with image size of 720×1168 (width×height) pixels and a bit depth of 8 bits per pixel in TIFF format [3].

In the pre-processing stage we cropped a ROI of 150×150 pixels (following the indications in [3]) using the information of the bounding box of the segmented region, preserving the aspect ratio, even when the lesion's dimensions are bigger than 150×150). When the lesion is next to the border of the image we translate the square crop, changing image coordinates and including the surrounding breast pattern, instead of zero-padding the outer portion of the crop. We have also performed data augmentation by using a combination of flipping and 90, 180 and 270 degrees' rotation transformations.

The networks used in this paper were previously used to perform classification in the ImageNet ILSVRC challenge data: the CNN-F (Fast, imagenet-vgg-f) model [5] and the Caffe reference model. The architecture of the CNN-F model consists in 8 learnable layers (5 convolutional layers and 3 fully-connected layers), and the fast processing is guaranteed by the 4 pixel stride in the first convolutional layer [5]. Caffe showed the best classification performance in our previous work and has a complete set of layers that are used for visual tasks such as classification and trains models by the fast and standard stochastic gradient descent algorithm [6]. In order to apply the pre-trained model to our problem, we have adapted the software MatConvNet [7] available for Matlab. Images were divided into 60% for training and 40% for testing, with an input size of 224×224 pixels (that is the size used for MatConvNet) and the parameters' exploration space comprised three fully connected layers, 50 epochs and five learning rate values (1e-2, 1e-3, 1e-4, 5e-2, 5e-3 and 5e-4).

## 4 Results and Discussion

Table 1 presents the results in terms of minimum and maximum of the images with the different methods of normalization. Note that in the Methods 1 and 3 the values range remains high, and in Methods 2, 4, 5 and 6 the range values are close to zero. Although, for example, the images *a*, *b* and *g* are visually similar, Table 1 shows that the minimum and maximum values are not the same.

Table 2 shows the results of normalization tests, performed five times, with the CNN-F and Caffe reference model in terms of area under the curve (AUC) mean and standard deviation and the statistic values ($p$ value) of comparison between the use or not of the different normalization methods. Note that only for Caffe and Method 2 the AUC value is statistically

[1] http://bcdr.inegi.up.pt

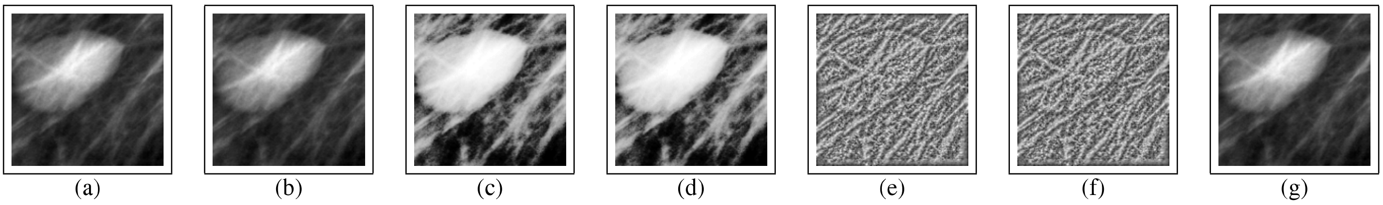


Figure 1: Examples of the same crop image (benign lesion) with the different normalization methods: (a-f) Methods 1 to 6 (g) NoNORM.

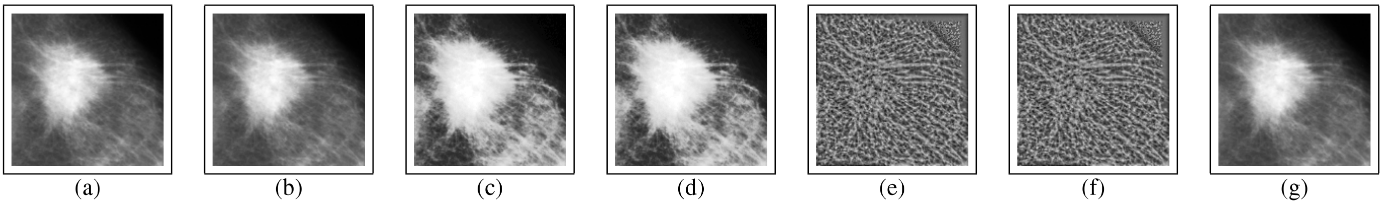


Figure 2: Examples of the same crop image (malign lesion) with the different normalization methods: (a-f) Methods 1 to 6 (g) NoNORM.

equal ($p$ value = 0.119), for the others always exist a significance difference for better or worse performance with more or less significance value. The best performance achieved with CNN-F was 0.786 and with Caffe 0.790 both using Method 2. However, note that with Caffe, the difference in the results between Method 2 and the one without normalization is not significant ($p$ value = 0.119), which leads to ask if whether important to make the image normalization. However, in the case of CNN-F, the AUC means with Method 2 reveal a significant improvement in the results, from 0.763 to 0.786 ($p$ value = 2.28e-5), with a best run of 0.793 against 0.768.

The histogram normalization does not seems to have a great influence in the performance of the network with Caffe; however with CNN-F, one observes an increase in the classification performance, although the results are slightly lower than those obtained with Method 2.

While GCN seems to have some effect in the network performance, mostly in CNN-F, the LCN does not produce any improvement in the results; the best AUC mean with these methods is 0.742. The AUC values of Methods 5 and 6 are similar, which may be due to the fact both methods have the same minimum and maximum values.

| Method | Benign lesion (Min/Max) | Malign lesion (Min/Max) |
|---|---|---|
| Met. 1 | -49.68 / 111.32 | -93.86 / 137.14 |
| Met. 2 | -1.52 / 3.41 | -1.82 / 2.66 |
| Met. 3 | -127.35 / 127.65 | -127.48 / 127.52 |
| Met. 4 | -1.70 / 1.70 | -1.70 / 1.70 |
| Met. 5 | -3.07 / 2.69 | -2.88 / 3.60 |
| Met. 6 | -3.07 / 2.69 | -2.88 / 3.60 |
| NoNORM | 82 / 243 | 4 / 235 |

Table 1: Examples of minimum and maximum values of benign and malign images, presented in figure 1 and 2, after the different normalization methods.

| Method | AUC Mean/Std<br>CNN-F | p-value | AUC Mean/Std<br>Caffe | p-value |
|---|---|---|---|---|
| Met. 1 | 0.767 / 0.003 | 0.0489 | 0.779 / 0.001 | 1.00e-6 |
| Met. 2 | 0.786 / 0.005 | 1.13e-4 | 0.790 / 0.002 | 0.119 |
| Met. 3 | 0.785 / 0.002 | 1.89e-6 | 0.781 / 0.003 | 1.89e-6 |
| Met. 4 | 0.785 / 0.002 | 1.08e-6 | 0.782 / 0.003 | 0.004 |
| Met. 5 | 0.730 / 0.003 | 1.17e-7 | 0.742 / 2.0e-4 | 1.03e-17 |
| Met. 6 | 0.729 / 0.003 | 4.37e-8 | 0.741 / 4.0e-4 | 2.17e-14 |
| NoNORM | 0.763 / 0.003 | - | 0.789 / 2.0e-4 | - |

Table 2: Results of normalization tests with the CNN-F and Caffe reference model (AUC mean and standard deviation) and statistic values of comparison between the use or not of the different normalization methods ($p$ value).

## 5 Conclusions

The effect of image normalization in the performance of the CNNs depends of which network is chosen to make the lesion classification. We have seen from the results that, for Caffe, the image normalization is not so important as much as for CNN-F. The method of image normalization that seems to have a bigger impact in the classification performance is the one that subtracts the image mean and divide by the standard deviation (Method 2). The use of LCN is associated with the worst results, which leads to believe that is not a good way to obtain a better CNN performance. The current study was made using scanned images; as future work, we intend to apply these methods to digital images, that are actually the most used in the medical field, with the aim of increasing the classification performance.

## 6 Acknowledgments

This work was supported by FCT - Fundação para a Ciência e a Tecnologia, through the UID/EEA/50008/2013 Project. The GTX Titan X used in this research was donated by the NVIDIA Corporation. The database used in this work was a courtesy of MA Guevara and coauthors, Breast Cancer Digital Repository Consortium.

## References

[1] Zhicheng Jiao, Xinbo Gao, Ying Wang, and Jie Li. A deep feature based framework for breast masses classification. *Neurocomputing*, 197:221–231, jul 2016.

[2] Rahimeh Rouhi, Mehdi Jafari, Shohreh Kasaei, and Peiman Keshavarzian. Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Systems with Applications*, 42(3):990–1002, 2015.

[3] John Arevalo, Fabio A González, Raúl Ramos-Pollán, Jose L Oliveira, and Miguel Angel Guevara Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 127:248–257, apr 2016.

[4] Ana Perre, Luis Alexandre, and Luis Freire. VI ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing, VipIMAGE. *Lesion Classification in Mammograms Using Convolutional Neural Networks and Transfer Learning*, Oct 18-20 2017.

[5] Simonyan K. Vedaldi A. Zisserman A. Chatfield, K. Best Scientific Paper Award Return of the Devil in the Details: Delving Deep into Convolutional Nets. *British Machine Vision Conference*, 2014.

[6] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[7] Lenc K. Vedaldi, A. MatConvNet – Convolutional Neural Networks for MATLAB. *Proceeding of the ACM Int. Conf. on Multimedia*, 2015.

# The Impact of Noise Removal on a Compression-based ECG Biometric Identification System

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

Several ECG (electrocardiographic) applications, both in the field of health, or even other subjects, like biometric identification, achieve better results when the signal is cleaned, i.e., without noise. In this paper, we have extended a method based in the *Hamming Distance*, that has proved able to find noise on an ECG signal. We study its effect in the accuracy results while performing ECG biometric identification, using a compression-based approach. We start by explaining how the method works, and then show the results on the experiments we performed, using real ECG data.

## 1 Introduction

ECG signals reflect an individual's cardiac electrical activity over a period of time. It has the advantage of being a unique aliveness indicator as it is difficult to be spoofed and falsified [3], which makes it desirable for biometric authentication purposes. However, this signal is prone to irregularities that are originated from several sources: pathological, psychological, noise, artifacts, among others [1, 2, 4, 5, 6, 11].

In a previous work [7], a method for finding noise has been proposed. However, it was only tested in synthetic ECG data. In this paper, we aim at extending that work, by using real data, and exploring the consequences of noise removal on a compression-based biometric identification system. The method is tested by using finite-context models (FCM) of different context depths $k$, as well as a mixture of FCMs.

### 1.1 Database

The database used in our experiments, and in previous works, was collected *in house* [4], where 25 participants were exposed to different external stimuli – *disgust*, *fear* and *neutral*. Data were collected on three different days (once per week), at the University of Aveiro, using a different stimulus per day.

The data signals were collected during 25 minutes on each day, giving a total of around 75 minutes of ECG signal per participant. Before being exposed to the stimuli, during the first 4 minutes of each data acquisition, the participants watched a movie with a beach sunset and an acoustic guitar soundtrack, and were instructed to try to relax as much as possible.

The ECG was sampled at 1000 Hz, using the MP100 system and the software AcqKnowledge (Biopac Systems, Inc.). During the preparation phase, the adhesive disposable Ag/AgCL-electrodes were fixed in the right hand, as well as in the right and left foot. We are aware that such an intrusive set-up is not desirable for a real biometric identification system. However, for testing purposes, it seems appropriate, as this approach is more reliable – produces less noise.

## 2 Method

### 2.1 R-peak detection

The development of a robust automatic *R-peak* detector is essential, but it is still a challenging task, due to irregular heart rates, various amplitude levels and *QRS* morphologies, as well as all kinds of noise and artifacts [9].

We have decided to use a *partially fiducial* method for segmenting the ECG signal and, since this was not the major focus of the work, we used a preexisting implementation to detect *R-peaks*, based on [9]. This method detects the *R-peak* by calculating the average point between the *Q* and *S* peaks (from the *QRS complex*) – this may not give the real local maximum of the *R-peak*, but it produces a very close point. Some evaluations were done using *R-peak* detection performed by humans, in order to validate this step.

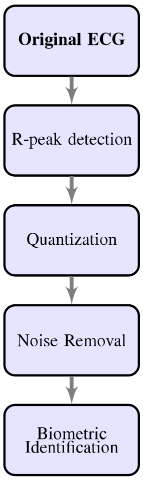


Figure 1: Overview of the different steps of the method proposed for biometric identification with noise removal.

### 2.2 Quantization

We consider that the signal is already discrete in the time domain, i.e., that it is already sampled. However, we perform re-sampling using the previously detected R-peaks.

The design of the *quantizer* has a significant impact on the amount of compression obtained and loss incurred in a lossy compression scheme. We have used the widely known Symbolic Aggregate ApproXimation [10], SAX, in order to quantize the ECG values into a discrete alphabet.

There is a fundamental trade-off to take into account while performing the choice of the *alphabet size*: the quality produced versus the amount of data necessary to represent the sequence. From previous experiments, we found that using an alphabet size of 6 and 200 symbols per each R-R segment (per "heartbeat") produced good results for biometric identification. However, this result does not guarantee that the same will hold true for a different dataset or application.

### 2.3 Noise Removal

The idea of our proposed approach is to store all the $n$ *words* (*SAX-words*) of an ECG on a size $n$ array, compute the $n-1$ Hamming-distances between those consecutive $n$ *words* and remove the *words* that correspond to an Hamming-distance greater than $\bar{x}+\delta\sigma$, were $\delta$ is a parameter of the method.

The distance $i$ is given by the distance of the word $i$ to the word $i+1$ (see Fig. 2) and, therefore, if it is greater than the threshold, both words $i$ and $i+1$ are removed.

### 2.4 Parameter Tuning

In order to tune the parameter $\delta$, we ran nearly 100 simulations, changing the parameter from 0.5 up to 3. Since the purpose of this experiment was only to tune the parameter, and not to obtain an optimal biometric identification accuracy, we used an extended-alphabet FCM based compressor (xaFCM) [8].

Because of this result, whenever we use "noise removal" on the results section, a value of $\delta = 1.1$ was used.

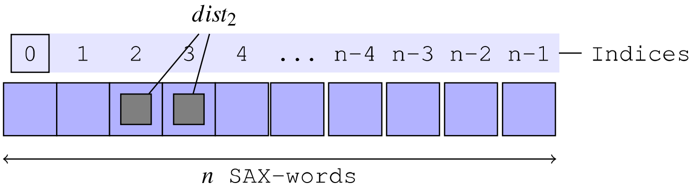


Figure 2: Array representation of SAX-words used to calculate Hamming-distances.

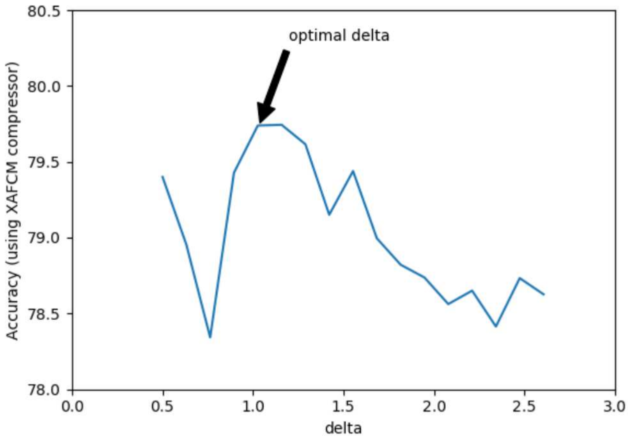


Figure 3: Parameter $\delta$ tuning.

## 3 Results

In order to simulate a real biometric identification system, whenever we test the system, we only use the other two days as reference (training). All tests were performed by using only 10 heartbeats as target (test).

From the results shown in Table 1, it is possible to notice that the noise removal slightly enhances biometric identification when using a single context FCM, specially when that context $k$ is low ($k = 10$ or $k = 13$). For high order FCMs, the noise removal does not seem to have much impact ($k = 20$) and, when using a mixture of FCMs, it even seems to be counter productive – which makes us think that these type of models are highly robust to noise, being able to capture some useful information, even when great amounts of noise are present.

## 4 Conclusions and Future Work

In order to improve ECG biometric identification, we plan on improving the compression of the ECG, by enhancing the mixture of FCMs to a "smart" one – this is possible to do by using a machine learning algorithm, like neural networks, on top of the compressors. We expect to do this in the near future and publish the results.

Table 1: Biometric identification accuracy using different sizes of finite-context models, with noise removal (NR) and without removal (WR). The Mixture used the contexts $k = 2, 4, 8, 12, 16, 20$ with $\alpha = 1, 1, 0.5, 0.1, 0.1, 0.001$, respectively, with a forgetting factor $= 0.99$.

| | Day for Target | | |
|---|---|---|---|
| **FCM context(s)** | Day 1 | Day 2 | Day 3 |
| $k = 10$ (WR) | 77.94% | 78.78% | 78.22% |
| $k = 10$ (NR) | 79.32% | 80.31% | 77.60% |
| $k = 13$ (WR) | 79.697% | 82.320% | 79.102% |
| $k = 13$ (NR) | 80.678% | 82.720% | 79.364% |
| $k = 16$ (WR) | 79.93% | 84.33% | 80.76% |
| $k = 16$ (NR) | 80.60% | 83.73% | 81.87% |
| $k = 20$ (WR) | 79.36% | 84.61% | 81.79% |
| $k = 20$ (NR) | 80.08% | 84.36% | 82.11% |
| Mixture (WR) | 82.45% | 84.98% | 83.13% |
| Mixture (NR) | 83.36% | 84.93% | 81.96% |

## 5 Acknowledgments

This work was partially supported by national funds through the FCT - Foundation for Science and Technology, and by european funds through FEDER, under the COMPETE 2020 and Portugal 2020 programs, in the context of the projects UID/CEC/00127/2013 and PTDC/EEI-SII/6608/2014. S. Brás acknowledges the Postdoc Grant from FCT, ref. SFRH/BPD/92342/2013.

## 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. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5999653`.

[2] Foteini Agrafioti and Dimitrios Hatzinakos. ECG biometric analysis in cardiac irregularity conditions. *Signal, Image and Video Processing*, 3(4):329–343, sep 2008. ISSN 1863-1703. doi: 10.1007/s11760-008-0073-4. URL `http://link.springer.com/10.1007/s11760-008-0073-4`.

[3] M Bassiouni and W Khalefa. A study on the Intelligent Techniques of the ECG-based Biometric Systems. *Recent Advances in Electrical Engineering*, 2015. URL `http://www.inase.org/library/2015/crete/COCI.pdf{\#}page=26`.

[4] 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. URL `http://www.ncbi.nlm.nih.gov/pubmed/26737619`.

[5] Susana Brás, Jacqueline Ferreira, Sandra C Soares, and Carlos F Silva. Psychophysiology of disgust: ECG noise entropy as a biomarker. In *37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pages 2351–2354, aug 2015. doi: 10.1109/EMBC.2015.7318865. URL `http://www.ncbi.nlm.nih.gov/pubmed/26736765`.

[6] Susana Brás, Sandra C Soares, Ricardo Moreira, and José M Fernandes. BeMonitored: Monitoring psychophysiology and behavior using Android in phobias. *Behavior research methods*, pages 1–9, jul 2015. ISSN 1554-3528. doi: 10.3758/s13428-015-0633-9. URL `http://www.ncbi.nlm.nih.gov/pubmed/26174715`.

[7] João M Carvalho, Armando J Pinho, and Susana Brás. Irregularity Detection in ECG signal using a semi-fiducial method. In *Proceedings of the 22nd RecPad*, pages 75–76, 2016.

[8] João M Carvalho, Susana Brás, Diogo Pratas, Sandra C. Soares, Jacqueline Ferreira, and Armando J. Pinho. Extended-Alphabet Finite-Context Models. *Pattern Recognition Letters (Submitted)*, 2017.

[9] P. Kathirvel, M. Sabarimalai, S. R. M. Prasanna, and K. P. Soman. An Efficient R-peak Detection Based on New Nonlinear Transformation and First-Order Gaussian Differentiator. *Cardiovascular Engineering and Technology*, 2(4):408–425, oct 2011. ISSN 1869-408X. doi: 10.1007/s13239-011-0065-3. URL `http://link.springer.com/article/10.1007/s13239-011-0065-3/fulltext.html`.

[10] J Lin, E Keogh, S Lonardi, and B Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *DMKD '03 Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11, 2003. URL `http://dl.acm.org/citation.cfm?id=882086`.

[11] Ikenna Odinaka, Po-Hsiang Lai, Alan D. Kaplan, Joseph A. O'Sullivan, Erik J. Sirevaag, and John W. Rohrbaugh. ECG Biometric Recognition: A Comparative Analysis. *IEEE Transactions on Information Forensics and Security*, 7(6):1812–1824, dec 2012. ISSN 1556-6013. doi: 10.1109/TIFS.2012.2215324. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6287579`.

# Poster Session III

# Active Learning GP-based Metamodels for Input Exploration in Transportation Simulation Models

Francisco Antunes
fnibau@uc.pt
Department of Civil Engineering
University of Coimbra
Coimbra, Portugal

Bernardete Ribeiro
bribeiro@dei.uc.pt
Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal

Francisco Pereira
camara@dtu.dk
DTU Management Engineering
Technical University of Denmark
Kongens-Lyngby, Denmark

## Abstract

Simulation modeling is a well-known and common approach to study real-world transportation systems, specially those that prove to be highly complex to be analyzed through conventional analytic methods. Many transportation policies are often designed and their performances assessed via simulation-based methodologies. However, such simulation models can become very time-consuming when detailed input domain exploration is needed. To tackle this problem, simulation metamodels are often used to approximate the simulator results.

In this paper, we propose an active learning algorithm based on Gaussian Processes (GP) metamodeling. Our algorithm gathers the most informative data points in batches, according to their variance and relative distance between them. This allows us to explore the simulator input space with fewer data points in a more efficient way, while avoiding computationally expensive simulation runs. We take advantage of the neighborhood notion encoded into the GP to select batches of points in such a way that they do not belong to the same high variance regions. In addition, we also suggest two simple and practical user-defined stopping criteria so that the iterative learning procedure can be fully automated. We illustrate our methodology with a simple example within a free and open-source microscopic simulator, which serves as the oracle in our active learning scheme.

## 1 Active Learning Metamodel Approach

Figure 1 depicts the general batch-mode active learning [4, 5] procedure used in this work and it serves as the base for other algorithms forwardly presented. Here, $\mathcal{L}$ represents the set of labeled training points, whereas $\mathcal{U}$ the set of unlabeled ones. The latter is defined according to input domain region we aim to explore. A GP [3] is used as our metamodeling approach [1]. In its essence, this algorithm selects batches of $k$ test points with highest variance (provided by the GP) in a way that each point does not originate from the same high variance region. Taking into account the spatial notion of closeness and similarity encoded into the GP via its kernel function, it is expected that spatially closer input points are more likely to have similar output values. Therefore, the hypothesis is that sampling multiple batch points from the same region is not efficient. To avoid this situation we introduce $\beta$ to thereby control the minimum distance between the selected active learning points. This parameter is a ratio w.r.t. the maximum possible distance between two any points (diameter) in the input space. Notice that this approach is only valid for continuous input metric spaces. Thus, if $\beta = 0.4$, then the minimum distance between the points is 40% of the maximum distance. We believe that it is more intuitive to provide a value in $[0,1]$ rather than choosing an absolute value arbitrary meaning. The parameter $k$ defines the size of the batch.

At each iteration a "new" GP model is fitted to $\mathcal{L}$, as the hyperparameters are obtained through the maximization of the likelihood function conditional on this training set. Then the trained GP is used to predict the simulation output values (labels) associated to the unlabeled points in $\mathcal{U}$, therefore avoiding many simulations runs. After, several ($k$) testing points are selected according to a given criteria, their respective true labels are obtained via oracle and finally $\mathcal{L}$ is expanded. This iterative process is repeated until the spotting criterion is satisfied.

We also propose two simple variance-based stopping criteria controlled by $\alpha$. The first, which we call Criterion A, states that the algorithm stops when the total current variance ($TCV$) w.r.t. the test points, at iteration $i$, is less than $(1-\alpha)\%$ of the initial total variance ($ITV$) at iteration 0. Thus, if, for example, $\alpha = 0.3$, the process stops when $TCV$ is reduced 70% w.r.t. to $ITV$, i.e., when $ITV(1-\alpha) \geq TCV$. Note that, in stead of the total variance, which is simply the sum of the variances of all test points, we could have considered the average variance per iteration. However, since our criterion is defined as a ratio, the total number of test point would cancel out. On the other hand, Criterion B is defined as a ratio between the average variance of the training ($AVTr$) points and the average variance of the test ($AVTs$) points at each iteration $i$. Here, contrary to Criterion A, since the total number of training and testing points is not the same, it makes sense to consider the average. When $AVTr/AVTs \geq \alpha$, the algorithm stops. The average variance at training points is less than the average variance at testing points, so this ratio lies in $[0,1]$. Moreover, as the process advances, $AVTs$ is likely to decrease, while $AVTr$ is expected to approximately maintain its values. However, this is a more demanding criterion to be satisfied if $\alpha$ is close to 1. If the model, in our case the GP, is very certain at the training points and the contrary at the test points, $AVTr/AVTs \approx 0$, which will prevent the algorithm from converging at an acceptable speed. Its performance will also depend on the noise structure of the underlying function being estimated.

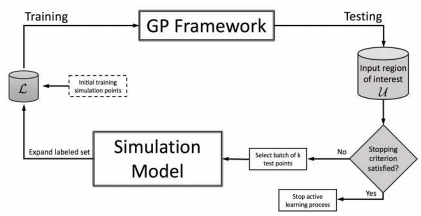

Training
GP Framework
Testing
$\mathcal{L}$
Initial training simulation points
Input region of interest $\mathcal{U}$
Expand labeled set
Simulation Model
Select batch of k test points
No
Stopping criterion satisfied?
Yes
Stop active learning process

Figure 1: Illustration of the proposed pool-based batch-mode active learning approach, with a simulation model playing the role of oracle.

Given the parametric nature of proposed base algorithm many variants can be derived depending on the concrete values assigned to $\alpha$, $\beta$ and $k$, as well as on the used stopping criterion. Due to space constrains, we only test the three following variations that are built upon the base structure of this general parametric: a) Algorithm 1: base algorithm + Criterion A, with no space restriction, i.e., $\beta = 0$; b) Algorithm 2: base algorithm + Criterion A, with space restriction, i.e., $\beta \geq 0$; c) Algorithm 3: base algorithm + Criterion B, also with space constrains. Algorithm 1 is the only approach lacking constrains over the input batch formation, so it represents our baseline batch-mode scheme.

## 2 Experiments

The studied example consists of a simple signalized intersection, depicted in Figure 2, from the Simulation of Urban Mobility (SUMO) [2] package. There is only traffic in three directions, North-South (NS), West-East (WE) and East-West. During each simulation run, the demand (traffic flow) generated from each operational axis is randomly generated according to a Poisson distribution, approximated by a Binomial distribution with parameter $p \in [0,1]$. This parameter sets how many vehicles are generated, on average, within a certain period of time. For example, if $p = 1/s$, then it means that one vehicle is expected every interval of $s$ seconds. In total, the simulated example encompasses three input parameters that have a direct influence in the intersection performance, namely, the NS, WE and EW demands, each of which associated with different

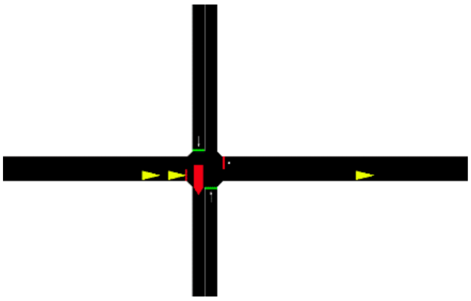
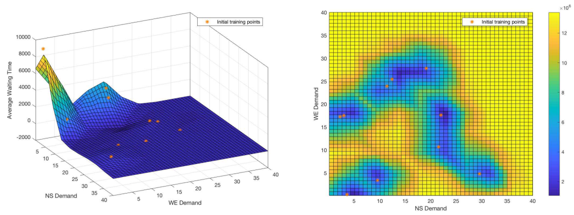
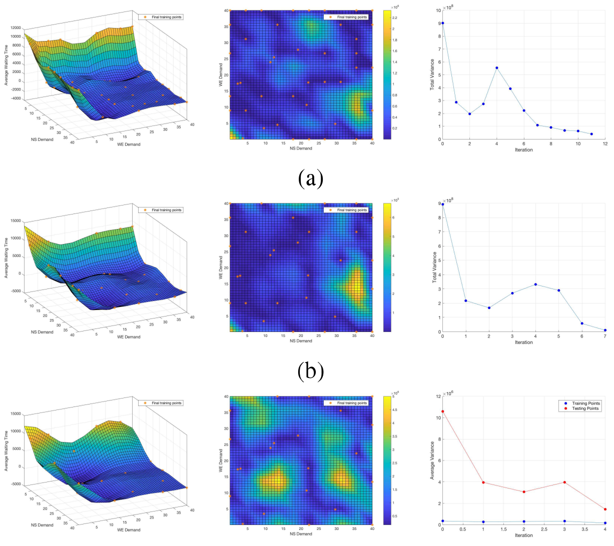


Figure 2: Visualization of the intersection with four approach lanes impleded in the SUMO example.



(a) (b)

Figure 3: Initial learning state for the traffic simulation data set for two inputs where the first training points correspond to 10 random simulation runs in input domain $[0,40]$, corresponding to Iteration 0. Panel (a) shows the first GP approximation surface and (b) the variance behavior across the input domain.



(a)

(b)

(c)

Figure 4: Final results for (a) Algorithm 1 - $\alpha = 0.95$, $\beta = 0$ and $k = 3$, (b) Algorithm 2 - $\alpha = 0.95$, $\beta = 0.3$ and $k = 3$ and (c) Algorithm 3 - $\alpha = 0.10$, $\beta = 0.3$ and $k = 3$.

Binomial parameters. We focus on the total waiting time spent by all the vehicles crossing the intersection as our aggregated traffic performance measure in terms of two simulation inputs: NS and WE axis demands. Our objective is to use our active learning approach to explore the input domain and to evaluate how it affects the total waiting time.

## 3 Results & Discussion

The input region of interest ($\mathcal{U}$) is defined by the square $[0,40] \times [0,40]$, from which, once again, 10 random points were selected, corresponding to the initial set of simulation runs ($\mathcal{L}$), as depicted in Figure 3. As expected, the variance near the training points is low when compared against the variance associated with the test points. Starting from this initial stage (Iteration 0), our approach is designed to actively search for the regions with highest variance, i.e., it preferably selects the next training batch within the yellow tones regions, as these are essentially characterized by a high degree of uncertainty. In any case, in this first approximation we can already observe that the values of the average waiting time (z-axis) are much higher when both NS and WE demands approach the origin.

The final results from Algorithms 1, 2 and 3 are respectively presented in panels (a), (b) and (c) from Figure 4. The set of experiments show that the GP approximations show identical simulation behaviors across the three active learning approaches. Suppose that the NS and WE demands are respectively defined by the following Binomial parameters $p_{\text{NS}} = 1/s_1$ and $p_{\text{WE}} = 1/s_2$. We can see that when both $s_1$ and $s_2$ tend to zero simultaneously, the average waiting time reaches its highest values, approximately between 1000 and 1400 seconds. The same observation is valid if one of demands is remained fixed. This is a common property shared by the three algorithms, which describe the simulator output behavior in terms of the studied input region. However, in terms of search efficiency and computational workload, the results show many significant differences. In Figures 4(a) and (b), we see that Algorithm 1 took 11 iterations to achieve a total variance reduction of 95% (Criterion A), against seven from Algorithm 2. Both are based on the same stopping criterion, however, due to the proposed space restriction ($\beta = 0.3$), the latter presents a more efficient performance, with a difference of $(11-7) \times 3 = 12$ simulation runs. This may be of little significance in the current academic example, but it can prove to be quite relevant in real-world and thus computational heavy simulations that can take up to days to terminate. Moreover, with regards to the final variance within the input region of interest, we can observe difference in the magnitude of the variance values. While Algorithm 2 presents an magnitude of $10^5$, whereas Algorithm 1 remained in the order of $10^6$. This shows that Algorithm 1 besides being less efficient, was also not able to equally reduce the total variance in comparison with Algorithm 2. Finally, Algorithm 3 attained interesting results fairly comparable to those of Algorithm 1. As seen in Figure 4(c), this approach only took four iterations to satisfy the Criterion B, i.e., it stopped when the average variance of the training points was approximately more than 30% of the variance average of the test set, since $\alpha = 0.3$. Although it may seem, at first glance, that Algorithm 3 performed better than Algorithm 2, it was actually less effective in terms of reducing the overall uncertainty across the test region.

## 4 Conclusion

Generally, by providing methodologies that efficiently explore the input spaces of simulation models, we are able to guide the simulation process towards the discovery of input-defined scenarios that represent alternative future system states. The identification of important inputs combinations that significantly affect the simulation outcome under analysis, is critical to build robust transport policies or even to assess where others have fail to meet their proposed goals.

The presented partial results originate from preliminary ongoing experiments. As future work, our goal is to effectively combine simulation metamodels with active learning strategies that not only provide new insights of the simulation model behavior, but also to enhance the decision making and taking processes, by searching for policy-relevant simulation input combinations. Bearing this objective in mind, we plan to develop more complex and robust active learning strategies that, along with simulation metamodeling approaches, are able to minimize the problem of expensive simulation runs.

## References

[1] Russell R Barton. Simulation metamodels. In *Simulation Conference Proceedings, 1998. Winter*, volume 1, pages 167–174. IEEE, 1998.

[2] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. Recent development and applications of sumo-simulation of urban mobility. *International Journal On Advances in Systems and Measurements*, 5(3&4):128–138, 2012.

[3] C. E. Rasmussen and C. Williams. *Gaussian processes for machine learning (Adaptive computation and machine learning)*. The MIT Press, 2005.

[4] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[5] Xizhao Wang and Junhai Zhai. *Learning from Uncertainty*. CRC Press, 2016.

# Improving Grasping Performance by Segmentation of Large Planar Surfaces

Vasco Lopes
http://www.di.ubi.pt
Luís A. Alexandre
http://di.ubi.pt/~lfbaa

Departamento de Informática
Universidade da Beira Interior
Instituto de Telecomunicações
6201-001 Covilhã, Portugal

## Abstract

Grasping objects is a task that humans do without major concerns. This results from learning and observing other skilled humans doing such task and with previous information, unconsciously, we know how to pick up different types of objects. However, grasping novel objects in unknown positions for a robot is a complex task which encounters many problems, such as the performance rates that are not perfect and the time consumption. In this paper we present a method that complements the state-of-the-art grasping by removing the largest planar surface of the image of the world before the grasp detector receives them. The proposed method improves the performance rate and is also capable of reducing the time consumption.

## 1 Introduction

Grasping novel objects is a very complex task for a robot and that's why it's an important area and with active and extensive research. In this paper we present a method to improve a current state-of-the-art algorithm that gives a robot the capability of grasping novel objects in unknown positions. Robots are getting more and more present in our daily basis, but some tasks still encounters many barriers, which is the case of grasping novel objects. The most predominant problems in the state-of-the-art methods are the incapacity of achieving perfect results in detecting grasps and the time spent on processing the algorithm for detecting such graps, because in a real-life situation, if the robot fails a grasp it can damage itself or persons that are around it, and if it spends too much time processing, the world can change and it executes movements that are not correct anymore and may collide with objects.

In a related work, Kehoe *et al.* [2] used the cloud to serve as a vast source of computation and data. The aggregation and sharing of training data proposed by this paper means that training multiple robots can occur faster than training on a single robot, which can be a way to address the problem of a robot encountering novel objects and having the cloud serving as a computation source can also decrease the run time.

Saxena *et al.* [5], presented a possible solution to the problem of grasping novel objects that the robot is perceiving for the first time thought vision. They propose a learning algorithm that doesn't require a 3D of the object, instead, the algorithm tries to identify a set of points in 2D images that corresponds to a good point at which to grasp the object, and with that point, it uses triangulation to obtain a 3D position to attempt the grasp.

Although there is work done in this area that tries to minimise the problems, there isn't a perfect solution. In this paper we propose an improvement to the method proposed in [1], which tries to calculate possible grasps by randomly selecting points from the point cloud and for each, calculating a surface normal and an axis of the major principle curvature of the object surface in the neighbourhood of that point. It generates potential hand candidates at regular orientations orthogonal to the curvature axis, and for each hand candidate it verifies if it is a possible grasp candidate and then classifies each one of the possible grasp candidates as either a viable grasp or not, using a deep neural network.

## 2 Proposed Method

The method proposed is an improvement to the original method described in [1]. In the original method, the grasp detector receives the information of the world directly, in our experiments using a Kinect. We propose a change to this method by introducing a segmentation node. This node receives the information of the world in form of a point cloud, calculates the largest planar surface using RANSAC, removes it from the image and sends it to the grasp detector node. The justification for removing the largest planar surface is that it usually represents a table top, the floor or a wall, and not the object to be grasped. By removing this large planar surface we are reducing the amount of data to be processed by the grasping algorithm. This can have two benefits: first, the potential grasps will not appear on the removed plane, increasing the probability that they are correctly placed on the object to be grasped. Second, since the potential grasps are more likely to be correct, the algorithm can work with a smaller number of attempts, to achieve the same grasping success rate, but using less time to do it. In figure 1, it's possible to see in a) the original image that is received by the segmentation node, in b) the plane, in red, that was calculated as being the largest one and in c) the segmented image without the plane in it.

The proposed method serves as an improvement in both time, because we are able to reduce the number of samples chosen to create grasp candidates, and performance in terms of the success in detecting viable grasps. A viable grasp is a grasp that a robot is able to perform. In figure 2 we can see two examples of where the grasp detector successfully detected viable grasps: they are indicated in the images as a blue parallel jaw gripper.

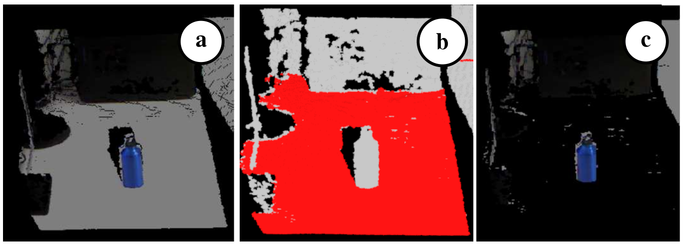


Figure 1: Three steps of the segmentation algorithm.

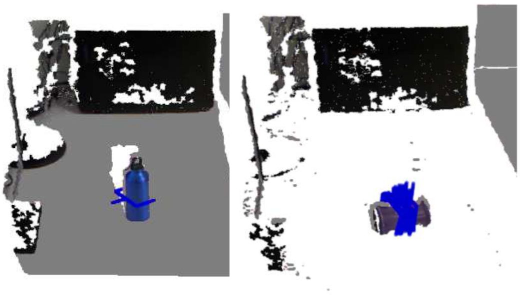


Figure 2: Two examples in which the grasp detection was successful. The left imagem shows a single grasping candidate and on the right image there are multiple succesfull candidates.

## 3 Experiments

We used the Robot Operating System (ROS) [3] and the Point Cloud Library (PCL) [4] for the implementation of this method. In order to compare the performance of the proposed method against the original one, we conducted three experiments, each one using the same set of 8 objects, which can be seen in figure 3. For each object we took 20 images with a Kinect in our lab, simulating a possible scenario where a robot needs to grasp an object on a table. Between each shot of an image we changed the object orientation and position so that each image differs from the others. After the images were captured, we segmented each one so we can evaluate the performance of the grasp detector with non-segmented images against segmented images. The average segmentation time was 0.21 seconds per image.

Each experiment had the same number of trials, 20 per object, using either original images (non-segmented) or segmented images. The grasps were not executed by a real robot: the method [1] identifies if a grasp is viable or not. A trial is classified as successful if the algorithm detects at least one viable grasp, if otherwise, the trial is classified as a failure.

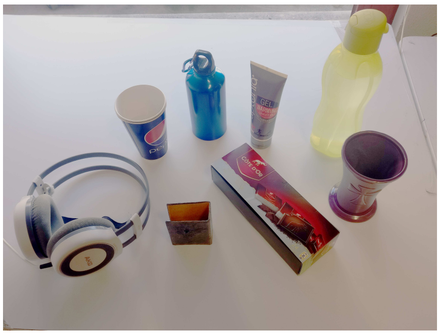


Figure 3: Set of 8 objects used to create the database.

### 3.1 Experiment 1

To serve as a baseline, we evaluated the performance of the original method with the original images (non-segmented). For this experiment we set the number of points to be sampled from the grasp generator, $n$, as 5000. The results of this experiment can be seen in table 1. The average run time was 2.63 seconds, this means that in each trial, the original method calculated the grasp candidates and classified them in 2.63 seconds, achieving a success rate of 45.63%, showing that the original method was capable of detecting a viable grasp in 73 of the 160 trials.

Table 1: Results of the experiment using non-segmented images with $n$=5000.

| | Success | Failure | Average run time (s) | Success Rate (%) |
|---|---|---|---|---|
| Blue Canteen | 19 | 1 | 2.65 | 95.0 |
| Cardboard Box | 11 | 9 | 2.69 | 55.0 |
| Cardboard Cup | 9 | 11 | 2.63 | 45.0 |
| Clay Cup | 11 | 9 | 2.66 | 55.0 |
| Gel Tube | 11 | 9 | 2.58 | 55.0 |
| Headphones | 6 | 14 | 2.64 | 30.0 |
| Paper Holder | 1 | 19 | 2.49 | 5.0 |
| Water Bottle | 5 | 15 | 2.69 | 25.0 |
| Overall | 73 | 87 | 2.63 | 45.63% |

### 3.2 Experiment 2

In this experiment, we tested the proposed method with the same number of sampling points as in 3.1, using the segmented images of the world. This experiment had as results a success rate of 71.88% and an average run time of 3.09 seconds, as can be seen in table 2. The total average run time of this experiment is 3.3 seconds. This is the sum of the average run time (3.09 seconds) and the average segmentation time for each image (0.21 seconds), which means that in this experiment, the proposed method is capable of segmenting an image, generate grasp candidates and classifying them with success in 115 of the 160 trials in an average run of 3.3 seconds.

Table 2: Results of the experiment using segmented images with $n$=5000.

| | Success | Failure | Average run time (s) | Success Rate (%) |
|---|---|---|---|---|
| Blue Canteen | 20 | 0 | 3.09 | 100.0 |
| Cardboard Box | 12 | 8 | 3.28 | 60.0 |
| Cardboard Cup | 19 | 1 | 3.08 | 95.0 |
| Clay Cup | 17 | 3 | 3.08 | 85.0 |
| Gel Tube | 18 | 2 | 3.17 | 90.0 |
| Headphones | 12 | 8 | 3.17 | 60.0 |
| Paper Holder | 8 | 12 | 2.94 | 40.0 |
| Water Bottle | 9 | 11 | 2.90 | 45.0 |
| Overall | 115 | 45 | 3.09 | 71.88% |

### 3.3 Experiment 3

In this experiment the proposed method was tested, with the same setup as the experiment described in 3.2 but with less sampling points. The number of points to be sampled in the image from the grasp detector in order to create grasp candidates, $n$, was reduced from 5000 to half, 2500. This experiment reduced the average run time of creating grasp candidates and classifying them to 1.62 seconds and achieved an success rate of 43.13% as described in table 3. This means that the total average run time is 1.83 seconds (1.62 seconds of the grasp detection run time and 0.21 seconds of segmenting the images) and that the grasp detector has able to detect viable grasps in 69 out of 160 trials.

Table 3: Results of the experiment using segmented images with $n$=2500.

| | Success | Failure | Average run time (s) | Success Rate (%) |
|---|---|---|---|---|
| Blue Canteen | 20 | 0 | 1.71 | 100.0 |
| Cardboard Box | 6 | 14 | 1.74 | 30.0 |
| Cardboard Cup | 10 | 10 | 1.60 | 50.0 |
| Clay Cup | 12 | 8 | 1.58 | 60.0 |
| Gel Tube | 12 | 8 | 1.64 | 60.0 |
| Headphones | 4 | 16 | 1.58 | 20.0 |
| Paper Holder | 1 | 19 | 1.51 | 5.0 |
| Water Bottle | 4 | 16 | 1.58 | 20.0 |
| Overall | 69 | 91 | 1.62 | 43.13% |

## 4 Conclusions

Grasping novel objects in unknown positions is a challenging task which doesn't have a perfect success rate and normally takes several seconds to process, which can be critical in a real-life scenario.

In this paper we present a method that can be used to determine a successful grasp faster or with a higher success rate, than a current state-of-the-art approach. It is capable of reducing the average run time in about 31% and still have similar success rate, achieving a result of 43.13% in 1.83 seconds, or, if the concern is not how fast the method runs but the success rate, the proposed method is able to increase the success rate in approximately 26%, from 45.63% to 71.88%, using a similar time as the original method.

Future work will consider the possibility of ataining both benefits simultaneously, shorter execution times and higher grasping performance, by optimizing the segmentation process to further reduce the amount of data to be processed.

## Acknowledgments

This work was supported by National Funding from the FCT - Fundação para a Ciência e a Tecnologia, through project UID/EEA/50008/2013.

## References

[1] Marcus Gualtieri, Andreas ten Pas, Kate Saenko, and Robert Platt. High precision grasp pose detection in dense clutter. *CoRR*, abs/1603.01564, 2016. URL `http://arxiv.org/abs/1603.01564`.

[2] B. Kehoe, A. Matsukawa, S. Candido, J. Kuffner, and K. Goldberg. Cloud-based robot grasping with the google object recognition engine. In *2013 IEEE International Conference on Robotics and Automation*, pages 4263–4270, 2013.

[3] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully B. Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. ROS: an open-source robot operating system. In *ICRA Workshop on Open Source Software*, 2009.

[4] R.B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 2011.

[5] Ashutosh Saxena, Justin Driemeyer, and Andrew Y. Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2):157–173, 2008.

# Land Cover Classification Competition

Borgine Gurué
up201502703@fe.up.pt
Ricardo Cruz
rpcruz@inesctec.pt
Kelwin Fernandes
kafc@inesctec.pt
Jaime S. Cardoso
jaime.cardoso@inesctec.pt

Faculty of Engineering, University of Porto (FEUP)
INESC TEC
Porto, Portugal

## Abstract

This report was prepared as the successful conclusion of the "Time Series Land Cover Classification Challenge" competition (TiSeLaC) in 3rd place out of 21 participants. The competition consisted of classifying patches of satellite data, among one of 9 distinct classes. For a few decades, satellite images have been used to observe and monitor the surface of Earth. In this challenge, the objective was to apply techniques of machine learning to classify Earth surface across a temporal series. The solution made use of a stacking cascade, using Nearest Neighbors and Random Forest as the underlying estimators.

## 1 Introduction

As part of the ECML-PKDD 2017 conference, the "Time Series Land Cover Classification Challenge" competition (TiSeLaC) [1] started July 1st, and lasted for 24 days. The testing set was only provided in the last four days to avoid cheating by means of manually labeling the data. Gurué participated in this competition as a follow-up to his masters thesis which was within the same domain.

The use of multispectral images or any other type of images to extract information has been the study objective of various areas of knowledge, such as medicine, agriculture, geology, and others. These images, by their complexity, provide us with features specific of the composition of a certain region of the Earth. In this challenge, the goal is to classify the surface of the land using multispectral images captured at various points in time. The area of study is the Reunion Island, France (see Figure 1).

Good classification of land surface is paramount to detect areas prone to fire and help avert various natural disasters. Also automatic classification can greatly aid urban and countryside planning teams on the ground.

## 2 Data

The dataset is the result of feature extraction based on 8 images acquired in 2014 above the Reunion Island (2866x2633 pixels at 30 m spatial resolution). Typically, photographs have only three channels (R, G and B). A multispectral image uses a wider spectral of bandwidths. These images usually have from 3 to 10 bands represented by pixels, and each band is acquired through a radiometer of remote detection. Usually, the definition of the classes of interest depends on the study field for which the images will be used to classify. A total of 10 features are provided; these are 7 surface band values, together with 3 radiometric indices (NDVI, NDWI and BI), computed from the first features [1].

NDVI is a composition of the redness bands, helping identifying vegetation. NDWI is a water index, composed of near-infrared and short-wave infrared wavelengths. BI is a brightness index. The computation of these features was not detailed in the competition, but these are typical features from the literature [4].

In the end, the dataset provided has 81,714 observations for training and 17,973 observations for evaluation, with 230 features: the 10 features detailed multiplied by 23 time points. Furthermore, the coordinates of each observation is also provided.

There were nine different classes to predict. Often, in these satellite imaging problems, the labels are not necessarily exclusive labels, i.e. an observation can have multiple labels. These are known as multi-labels. But this competition was a classic classification problem, where the labels are exclusive. The nine classes proposed in the challenge are Urban Areas (20%), Forests (20%), Sparse Vegetation (20%), Rocks and bare soil (16%), Sugarcane crops (9%), Grasslands (7%), Water, Other crops, and Other built-up Surfaces. The latter with less than 5% of the data.

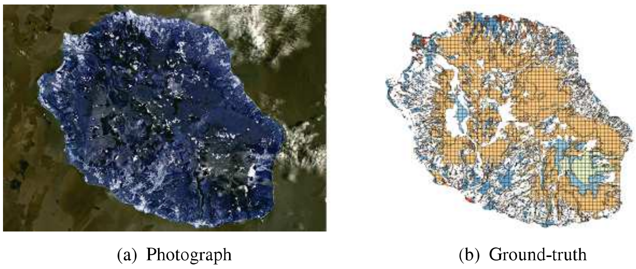

(a) Photograph (b) Ground-truth

Figure 1: The Reunion Island site.

## 3 Evaluation

The evaluation of the competition was a weighted $F_1$ score. $F_1$ score is defined for a binary problem as

$$F_1 = \frac{2TP}{2TP+FN+FP},$$

where TP, FN and FP are the True Positives, False Negatives and False Positives, respectively. In this challenge, since the predicted variable is multi-label, a weighted average was used, which was called F-metric,

$$\text{F-metric} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}F_1(k),$$

where $F_1(k)$ is the $F_1$ score considering label $k$ as positive and all others as negative, $n_k$ is the number of samples of label $k$, and $K$ are the number of labels.

## 4 Model

As mentioned in the previous sections, we propose to use two base models: Nearest Neighbors and Random Forest, and then combining them using an ensemble technique known as cascade stacking. In the following sections, we describe the models and technique used.

### 4.1 Nearest Neighbors

The $k$-Nearest Neighbors ($k$NN) is a non-parametric model; that is, it makes no assumptions about the data. It is an instance learning model, where $k$ is the number of closest instances against any new instances are evaluated against. Usually, Euclidean distance is used to find the closest neighbors, as we used. Majority voting among the $k$ neighbors is then used to choose the class of the new instance.

$k$NN has a particularity in that, as the amount of data approaches infinity, it approaches Bayes error. Furthermore, it has some lower bounds, such as in the two-class case, where the algorithm is guaranteed to yield an error rate no worse than twice the Bayes error rate [3]. Given the amount of training data provided in this challenge (81,714 observations), we decided to make use of this model as a base estimator. The (latitude, longitude) pairs were used as the exogenous variables.

We have used $k$NN to classify observations using only the secondary coordinate dataset that was also provided in the competition. A grid search hyperparametrization was performed; the model was found to be very resilient to $k$ in this dataset, and $k$=3 was finally used.

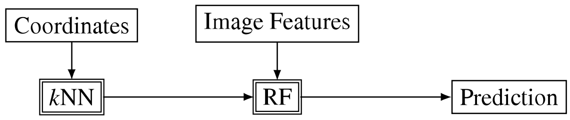


Figure 2: The stacking approach used in the challenge (model in double border; data in single border).

### 4.2 Random Forest

Random Forest (RF) is a combination of decision trees, such that each decision tree is trained in a different sampling of the original data by bootstrapping both features and observations, therefore keeping the same data distribution. Single decision trees often have high variance, and RF greatly ameliorates that, building a robust classifier [2]. As in *k*NN, majority voting is used to decide the class of new instance, for classification problems. RF is widely used in this problems of multispectral classification [5, 6].

The number of decision trees used in our RF was of 1,000 to ensure a robust model, while keeping training time sensible. Yet, training time took almost 15 minutes.

### 4.3 Stacking

Stacking is the usage of multiple models in order to compute the final result. The idea is to make use of different features of each model, or to use different models in different feature spaces. Usually, a weighted average or a different model is used to compute the final prediction. In the particular case of this challenge, we have first use the *k*NN model using only the coordinates to find the class of the land surface near the new instance, and then used the *k*NN prediction as a new feature for the RF. The RF used the 230 features based on the image together with the predictions of the *k*NN. See Figure 2 for a diagram of the proposed solution.

### 4.4 Cross-validation

Two cross-validation approaches were considered. The first approach, hereby termed *spatial*, consisted in splitting the data into non-overlapping geographical training and testing sets; data is divided into four geographical quadrants thus producing four folds. The other cross-validation approach, termed *random*, is a typical random split using stratified *k*-fold, with $k$=4.

Typically, evaluation data is split spatially to ensure no correlation between the training labels and the final labels, otherwise the model can simply memorize predictions associated to the spatial patterns. However, we were unsure what strategy the organizers would be used to produce the evaluation data.

## 5 Results

Stacking was not found to be optimal in either validation case, but we were unsure what strategy the organizers would be used to produce the evaluation data, and stacking seemed to have found a good compromise between each solution. When splitting data *spatially*, *k*NN was naturally subpar, since the closest training neighbors were too distant of the testing observations. But, based on the distribution of the training coordinates, we suspected organizers would use a random validation strategy, in which case *k*NN produced above pair results.

The final model produced a final F-measure score of 53% and 96% for the spatial CV and random CV cases, respectively. Fortunately, random CV was used by the organizers as this score was close to the final 97% score of the leaderboard (see Table 2).

The most important features, computed using feature importance [2], were the *k*NN predictions (17%), NVDI (17%), the 5th band (14%), and NDWI (10%), with the rest below 10%. In terms of time, *k*NN, took less than one second for both training and testing, while RF took an average of almost 10 minutes for training, and less than one second for testing.

| | Validation | |
|---|---|---|
| Model | Spatial | Random |
| *k*NN | 21.92 ± 4.6 | 97.75 ± 0.1 |
| RF | 68.18 ± 10.8 | 88.53 ± 0.2 |
| Stack | 53.29 ± 10.5 | 96.16 ± 0.1 |

Table 1: F-metric results along the stack for two validation approaches (higher=better).

| Place | Team | F-metric |
|---|---|---|
| 1st | baML | 99.29 |
| 2nd | COMANDA | 99.03 |
| 3rd | BG | 96.97 |
| 4th | TeamTUM | 96.32 |
| ... | ... | ... |
| 20th | UFMG_PATREO | 74.18 |
| 21th | TWDTW | 63.11 |

Table 2: The TiSeLaC challenge results.

## 6 Conclusion and Future Work

Land classification using multispectral satellite images is of great importance to prevent fires and other natural disasters, and to miscellaneous other uses. In this competition, the challenge was to automatically classify Reunion Island (France) landscape. Our approach involved a cascade stacking, where the output of *k*-Nearest Neighbors using only coordinates was fed into the Random Forest which also used the various image features. The organizers produced the evaluation set using a random strategy, which we exploited, ending up at 3rd place in the competition with a score of 96.97%.

As future work, we could partially rebuild the initial satellite image by combining the multispectral features with the coordinate from which each feature was taken. A convolutional neural network could then be used using this image as input for the classification.

## Acknowledgments

This work was funded by the Project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF), and also by Fundação para a Ciência e a Tecnologia (FCT) within PhD grant numbers SFRH/BD/122248/2016 and SFRH/BD/93012/2013.

## References

[1] TiSeLaC: Time series land cover classification challenge. https://sites.google.com/site/dinoienco/tiselc. Accessed: 2017-08-12.

[2] L Friedman Breiman, JH Friedman, Stone Olshen, and C Stone. Cj, 1984. classification and regression trees. *Pacific Grove, Kalifornien*, 1983.

[3] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[4] Tiago MA Santos, André Mora, Rita A Ribeiro, and João MN Silva. Fuzzy-fusion approach for land cover classification. In *Intelligent Engineering Systems (INES), 2016 IEEE 20th Jubilee International Conference on*, pages 177–182. IEEE, 2016.

[5] Shaohong Tian, Xianfeng Zhang, Jie Tian, and Quan Sun. Random forest classification of wetland landcovers from multi-sensor data in the arid region of xinjiang, china. *Remote Sensing*, 8(11):954, 2016.

[6] Konrad J Wessels, Frans van den Bergh, David P Roy, Brian P Salmon, Karen C Steenkamp, Bryan MacAlister, Derick Swanepoel, and Debbie Jewitt. Rapid land cover map updates using change detection and robust random forest classifiers. *Remote Sensing*, 8(11): 888, 2016.

# Indoor Environment and Human Shape Detection for Augmented Reality: an initial study

Veiga, Ricardo J.M.
ricardojorge@martinsveiga.com
Bajireanu, Roman
romanbajireanu@gmail.com
Pereira, João A.R.
japereira@ualg.pt
Sardo, João D.P.
joao_dps@outlook.com
Cardoso, Pedro J.S.
http://w3.ualg.pt/~pcardoso/
Rodrigues, João M.F.
http://w3.ualg.pt/~jrodrig/

LARSyS & Instituto Superior de Engenharia
University of the Algarve
Faro, Portugal

## Abstract

The application of the most recent technologies is fundamental to improve user museum experience. One of those technologies is Augmented Reality. With the objective of developing a mobile integrated Augmented Reality framework, this paper presents the initial study to develop two sub-modules of this framework: the indoor environment detection and human shape detection. The final goal is to use mobile devices with RGB cameras to on-the-fly detect the environment (e.g., wall, ceiling, floor) and humans, in order to project over them contents (e.g., pictures, videos) and other information related to the museum pieces. Initial results show that despite these sub-modules being still in embryonic state, they already show positive feedbacks.

## 1 Introduction

Augmented Reality (AR) [2] is a technology that, thanks to the mobile devices increasing hardware capabilities, quickly evolved in the recent years, gaining a huge amount of users. AR empowers a higher level of interaction between the user and real world objects, extending the experience on how the user sees and feels those objects. The M5SAR: Mobile Five Senses Augmented Reality System for Museums project [11] aims for the development of an AR system to be a guide in cultural, historical and museum events. This is not something new, since almost every known museum has its own mobile applications (App) [15], although Apps with AR are more uncommon, see e.g. [6]. The novelty of the M5SAR project is to extend the AR to the human five senses, see e.g. [11].

One of M5SAR's modules is MIRAR - Mobile Image Recognition Augmented Reality framework. MIRAR [11] focus on the development of a mobile multi-platform AR framework, with the following main goals: (a) perform "all" computational processing in the client-side (mobile device); (b) use real world two- and three-dimensional (2D and 3D) objects as markers for the AR; (c) allow to project contents (e.g. media) on the objects displayed in the mobile device's screen, as well as enhance the object's displayed contents, by touching regions on the device's screen; (d) recognize environments, allowing the projection of contents on them; (e) detect human shapes and allow projection of contents on those shapes; (f) use the mobile device's RGB camera to achieve these goals. A framework that integrates all these goals is completely different from the existing SDKs, frameworks, content management AR systems, see e.g. [1]. In this paper we focus on the initial sub-modules for recognizing environments and human shapes using the mobile RGB camera (d-f).

There is an increased necessity for environment detection, from AR to robotics. The usual approach for image acquisition involves the use of RGB-D devices or LIDAR sensors [9]. Man-made environments (such as museums) are characterized by the existence of numerous parallel lines and orthogonal edges from which the geometry of any said space can be retrieved [10, 14]. As a M5SAR project restriction, considering that the input will come from mobile (RGB) cameras, all the environments will be captured from a perspective view, where all the receding lines of said perspective will converge into a unique point in the horizon that is called the vanishing point [10, 14]. From there it is possible to determine the different planes (e.g., wall, floor, ceiling) of the environment, discarding irrelevant information [10]. In parallel, recent years have witnessed significant progress in object detection due to the use of Convolutional Neural Networks (CNNs). There are many recent object detectors based on these networks, such as Single Shot Multibox Detector (SSD) [8], You Only Look Once (YOLO) [12], DeepMultiBox [5], or Faster R-CNN [13], with some of them fast enough to be run on mobile devices and good enough to detect humans in real time.

In the following sections we present the initial algorithm for mobile environment recognition and projection and the initial tests for human shape detection. It is important to stress, due to M5SAR restrictions, all software has to be implemented using Unity 3D [16], consequently using the OpenCV [4] asset for Unity.

## 2 Environment detection and projection

As mentioned above, the goal of this MIRAR sub-module is the detection of environments, in this case walls, through RGB images (frames acquired from a mobile device), and the superimposition of contents, such as video or images, on the detected walls. The major steps of the algorithm are as follows: (a) read the input frame; (b) convert it to CIELAB (perceptual uniform color space); (c) apply Canny edge detection [3]; (d) apply the probabilist Hough transform [7]; (e) for each pair of points returned, the lines are retrieved as $y = m_i x + b_i$, and average of the ones that share a $m_i$ and $b_i$ with a 5% deviation; (f) horizontal, vertical and vanishing lines are differentiated and the vanishing point is found; (g) ideal lines are selected using a color comparison method of patches above/left and below/right each line; (h) calculate all intersection points between lines; (i) select and order the corner points (for each plane); (j) apply perspective wrapping to a content with the corners points; (k) replace the new warped pixels in the original frame for each plane.

The first steps (a-e) consist in transform the input image (Fig. 1, top row left) from RGB to CIELAB, apply the Canny edge detection to the L channel (Fig. 1, top row right) and the Hough transform to detect the lines, which are averaged over the detected lines with a 5% deviance; this allows the elimination of overlapping lines. All remaining lines are identified and flagged (Fig. 1, middle row left). After that, steps (f-i), the vanishing lines are analysed and the vanishing point is found. The lines that do not intersect at the vanishing point and are neither (quasi-)horizontal or (quasi-)vertical are discarded (Fig. 1, middle row right). The remaining vanishing lines are identified for each side of the vanishing point. Afterwards, random patches above and below the vanishing lines are color compared using the CIELAB color space to identify the pairs with significant variation, as these are expected to be limiting the environment's walls. The points intersecting with the horizontal lines are then retrieved, and the vertical lines within are discarded. The intersection with the remaining vertical lines is then obtained. The perspective points of $y$ when $x$ is at its minimum and at its maximum width are also collected. This way, even if points are found outside of the frame, the perspective of the content in the plane is maintained. Next, a left to right sweep is performed allowing to segment how many vertical planes (in this context, the walls) there are in the environment. Groups of points are then delivered for each plane, see Fig. 1, 3rd row left.

After obtaining the coordinates from the plane for where the content projection is intended, it is performed an evaluation of how much from the image to be projected is necessary to fill in the planes of the screen. Each

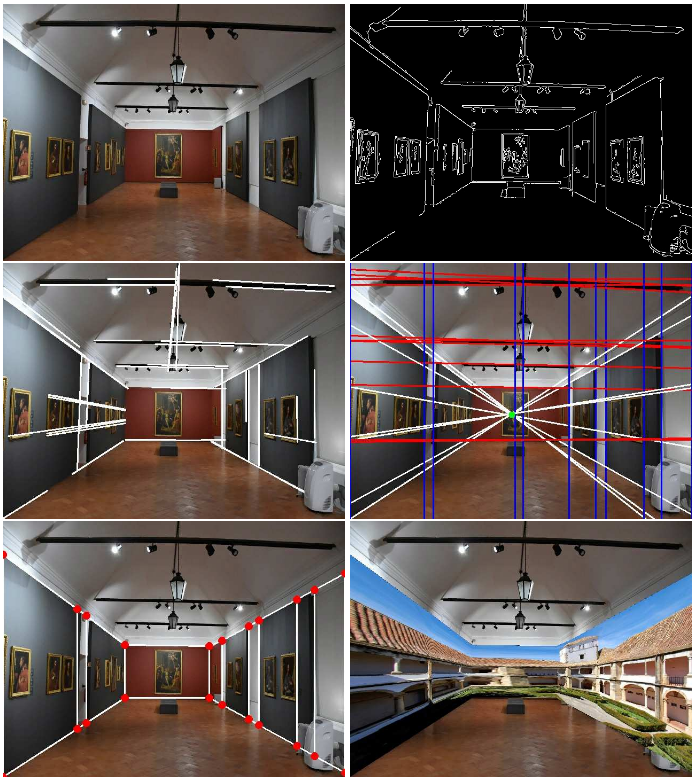


Figure 1: Left to right, top row, museum image and Canny Edge Detector, middle row, Hough Transform and horizontal, vertical and vanishing lines. Bottom row, intersecting points and warped image superimposed.

plane is processed at a time, from left to right. Using these coordinates together with those from the original width of the image, it is obtained the transform from the perspective, which is then applied over the content to be projected (Fig. 1, 3rd row right).

## 3 Human shapes detection

For the human shapes detection three models were tested in the mobile device, SSD-Mobilenet [8], YOLO [12], and DeepMultiBox [5]. Empirical tests were done using a ASUS Zenpad 3S 10 (see Fig. 2), showing that in real world conditions the SSD-Mobilenet presented a better accuracy and speed (validating what is mentioned in [8]) than the other two models; nevertheless more consistent tests need to be performed in different museum conditions. It is important to stress at this point that the detection human shapes will allow the projection of contents on those shapes/persons, as for instance "to dress" the museum users with the clothes corresponding to the object epoch that is being detected by the AR.

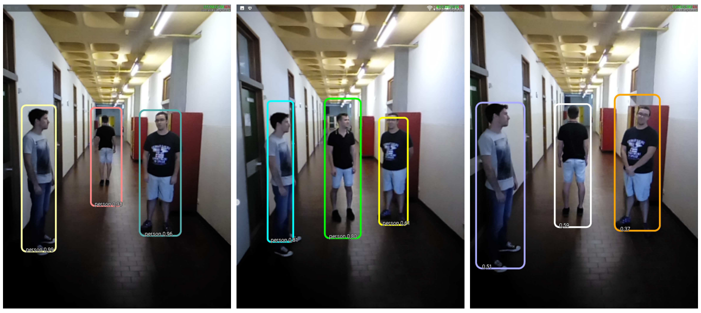


Figure 2: Left to right, examples of human detection with SSD-Mobilenet, YOLO and DeepMultiBox.

## 4 Conclusions

This paper presented the initial algorithm and tests performed for mobile environment recognition and projection, as well as presented the initial models tested for human shape detection. Despite being in the embryonic stage of integration in the MIRAR's framework, both show positives feedbacks. Nevertheless, the environment recognition algorithm has to be optimized to achieve "on-the-fly" performance, using the information of previous frames (detected walls) as well as the image flow information. In the case of the human detection models, they present very positive initial results, nevertheless, the final goal of projecting information (e.g., images and video) over the human body is yet to be implemented.

Future work will also consist in the creation of an adaptive environment detention, with its classification supported in the vanish point. In the case of human shape detection, the goal to segment the person in head, torso, harms, legs, etc. in a way to apply projected (AR) clothes.

## Acknowledgments

This work was supported by the projects: FCT LARSyS UID/EEA/50009/2013, and M5SAR I&DT nr. 3322 financed by CRESC ALGARVE2020, PORTUGAL2020 and FEDER. We also thank Faro Municipal Museum and the M5SAR project leader, SPIC - Creative Solutions [www.spic.pt].

## References

[1] Artoolkit. ARtoolKit, the world's most widely used tracking library for augmented reality. `http://artoolkit.org/`, 2016. Retrieved: Nov. 16, 2016.

[2] Ronald Azuma, Yohan Baillot, Reinhold Behringer, Steven Feiner, Simon Julier, and Blair MacIntyre. Recent advances in augmented reality. *IEEE Computer Graphics and Applications*, 21(6):34–47, 2001.

[3] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[4] Enox. OpenCV for Unity. `https://goo.gl/MuxFj2`, 2017. Retrieved: April 04, 2017.

[5] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2154, 2014.

[6] HMS. Srbija 1914 / augmented reality exhibition at historical museum of Serbia, Belgrade. `https://vimeo.com/126699550`, 2017. Retrieved: April 04, 2017.

[7] Paul VC Hough. Method and means for recognizing complex patterns. Technical report, 1962.

[8] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. *arXiv preprint arXiv:1611.10012*, 2016.

[9] Rostislav Hulik, Michal Spanel, Pavel Smrz, and Zdenek Materna. Continuous plane detection in point-cloud data based on 3D Hough transform. *Journal of Visual Communication and Image Representation*, 25(1):86–97, 2014.

[10] M. Moreno, Somayeh Shahrabadi, João José, J.M. Hans du Buf, and João M.F. Rodrigues. Realtime local navigation for the blind: detection of lateral doors and sound interface. *Procedia Computer Science*, 14:74–82, 2012.

[11] J.A.R. Pereira, J.D.P. Sardo, M.A.G. Freitas, Veiga R., P.J.S. Cardoso, and J.M.F Rodrigues. Mirar: Mobile image recognition based augmented reality framework. *Int. Congress on Engineering and Sustainability in the XXI Century*, 2017.

[12] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2017.

[14] M Serrão, Somayeh Shahrabadi, M Moreno, JT José, José I Rodrigues, João MF Rodrigues, and JM Hans du Buf. Computer vision and gis for the navigation of blind persons in buildings. *Universal Access in the Information Society*, 14(1):67–80, 2015.

[15] TWSJ. The wall street journal: Best apps for visiting museums. `https://goo.gl/cPTyP9`, 2017. Retrieved: April 04, 2017.

[16] Unity. Unity3D. `https://unity3d.com/pt`, 2014. Retrieved: Nov. 10, 2014.

# The Potential of Multimodal Learning for Sign Language Recognition

Pedro M. Ferreira [1,2]
pmmf@inesctec.pt
Jaime S. Cardoso [1,2]
jaime.cardoso@inesctec.pt
Ana Rebelo [1]
arebelo@inesctec.pt

[1] INESC TEC
Porto, Portugal
[2] Universidade do Porto
Porto, Portugal

## Abstract

Automatic Sign Language Recognition (SLR) has becoming one of the most important research topics in the field of human computer interaction. The purpose of such systems is to automatically translate the signs from video or images into text or speech, attempting to tear down the communicational barrier between deaf and hearing people. The aim of this paper is to exploit multimodal learning techniques for an accurate SLR, making use of data provided by Kinect and Leap Motion. In this regard, single-modality approaches as well as different multimodal methods, mainly based on convolutional neural networks, are proposed. Experimental results demonstrate that multimodal learning yields an overall improvement in the recognition accuracy.

## 1 Introduction

Sign language (SL) is an integral form of communication especially used by hearing impaired people within deaf communities worldwide. It is a visual means of communication, with its own lexicon and grammar, that combines articulated hand gestures along with facial expressions to convey meaning. As most of hearing people are unfamiliar with SL, deaf people find it difficult to interact with the hearing majority. In this regard, Sign Language Recognition (SLR) has becoming an appealing topic in modern societies. Its main purpose is to automatically translate the signs from video or images into the corresponding text or speech.

The SLR task can be addressed by using wearable devices or vision-based approaches. Vision-based SLR is less invasive since there is no need to wear cumbersome devices that might affect the natural signing movement. The SLR problem was first addressed by the computer vision community by means of just using the colour information of images and videos [1]. Afterwards, several works tried to combine colour and depth information, making use of low-cost consumer depth cameras (e.g., Microsoft's Kinect), for a more accurate sign recognition [2]. More recently, the introduction of the Leap Motion has launched new research lines for gesture recognition. Instead of a complete depth map, the Leap Motion sensor directly provides the 3D spatial positions of the fingertips and the hand orientation with quite accuracy ($\approx 200\mu m$). Although Leap Motion may have a great potential for sign recognition, it is not always able to recognize all fingers in some hand configurations. In order to overcome that limitation, Marin *et al.* [3, 4] combined the input data from Leap Motion with Kinect.

In this work, we extent the ideas proposed in [3] and [4], improving their results. In particular, our main contributions are:

- the development of a multimodal learning framework for the SLR, making use of data provided by both Kinect and Leap Motion.
- a comprehensive comparison between single-modality and multi-modal learning techniques, in order to demonstrate their effectiveness in the overall sign recognition performance.

## 2 Methodology

The aim of this paper is to explore the potential of multimodal learning for SLR. To accomplish this purpose, single-modality approaches as well as different multimodal methods, to fuse them at different levels, are proposed. Multimodal techniques include feature-level and decision-level fusion methods.

### 2.1 Single-modality Sign Recognition

In this section, the implemented single-modality methodologies for SLR are presented. For both kinect modalities (colour and depth), we resorted to a deep learning strategy based on convolutional neural networks (CNNs); whereas for Leap Motion we implemented a traditional machine learning pipeline with hand-crafted feature extraction.

#### 2.1.1 Kinect modalities (colour and depth)

The implemented neural network follows the traditional CNN architecture for classification, typically starting from several sequences of convolution-pooling layers to fully connected layers. Hence, the implemented CNN is composed by two convolution layers and one fully connected layer (or dense layer), in which each convolution layer is followed by a $2 \times 2$ max-pooling layer. Both convolution layers have the same filters' number and size (32 and 5, respectively). Finally, the last layer of the CNN is a soft-max output layer, which contains the output probabilities for each class label. The Nesterov's Accelerated Gradient Descent with momentum was used for optimization, and the categorical cross-entropy was used as the loss function. During the training stage, several regularization techniques were applied to prevent overfitting (i.e., dropout and data augmentation).

#### 2.1.2 Leap Motion

Unlike Kinect, Leap Motion does not provide a complete depth map, instead it directly provides a set of relevant features of hands and fingertips. From these data, 3 different types of features were computed:

1. **Fingertip distances** $D_i = \|F_i - C\|, i = 1,...,N$; where $N$ denotes the number of detected fingers and $D_i$ represents the 3D distances between each fingertip $F_i$ and the hand centre $C$.
2. **Fingertip inter-distances** $I_i = \|F_i - F_{i+1}\|, i = 1,\ldots,N-1$; represent the 3D distances between consecutive fingertips.
3. **Hand direction** $O$: represents the direction from the palm position toward the fingers. The direction is expressed as a unit vector pointing in the same direction as the directed line from the palm position to the fingers.

Both distance features are normalized by signer (user), according to the maximum fingertip distance and fingertip inter-distance of each user. Then, these 3 features are used as input into a multi-class SVM classifier for sign recognition.

### 2.2 Multimodal Sign Recognition

The data provided by Kinect and Leap Motion have quite complementary characteristics, since while Leap Motion provides few accurate and relevant keypoints, Kinect produces both a colour image and a complete depth map with a large number of less accurate 3D points. Therefore, we intend to exploit them together for SLR purposes.

According to the level of fusion, multimodal fusion techniques can be roughly grouped into two main categories: (i) feature-level, and (ii) decision-level fusion techniques [**?** ]. As described in the following, we propose multimodal approaches of each fusion category for the SLR task, making use of colour, depth and Leap Motion data.

#### 2.2.1 Feature-level fusion

In general, feature-level fusion is characterized by three phases: (i) learning a representation, (ii) supervised training, and (iii) testing. According to the order in which phases (i) and (ii) are made, feature-level fusion

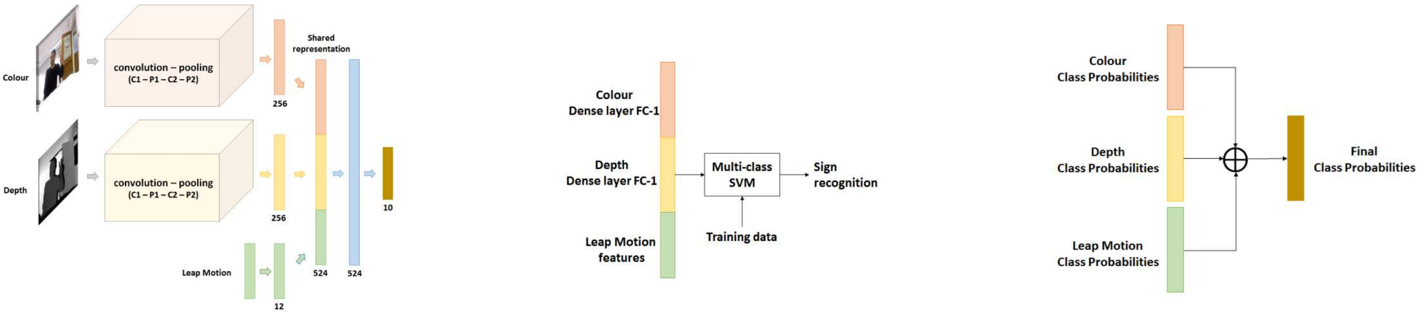

Colour
convolution – pooling (C1 – P1 – C2 – P2)
Shared representation
256
Depth
convolution – pooling (C1 – P1 – C2 – P2)
256
10
Leap Motion
12
524
524
Colour Dense layer FC-1
Depth Dense layer FC-1
Leap Motion features
Multi-class SVM
Sign recognition
Training data
Colour Class Probabilities
Depth Class Probabilities
Leap Motion Class Probabilities
⊕
Final Class Probabilities

Figure 1: Multimodal sign recognition: End-to-end feature-level fusion (left), Multi-step feature-level fusion (middle) and Decision-level fusion, where ⊕ is an aggregate operator representing the decision rule for fusion (right).

Table 1: Results of the proposed single-modality and multimodal recognition methodologies in terms of classification accuracy (%). C, D and L denote colour, depth and leap motion data, respectively.

| Method | Involved modalities | Acc (%) |
|---|---|---|
| *Single-modality methods* | | |
| **CNN** | C | 93.50 |
| **CNN** | D | 91.75 |
| **Hand-crafted** | L | 82.00 |
| *Multimodal methods* | | |
| ***feat_end2end*** | C+D | 93.00 |
| | C+D+L | 94.25 |
| ***feat_mstep*** | C+D | 96.25 |
| | C+D+L | 96.75 |
| ***late_mean*** | C+D | 96.00 |
| | C+D+L | **97.00** |
| ***late_conf*** | C+D | 96.00 |
| | C+D+L | 96.50 |
| ***late_learn*** | C+D | 96.25 |
| | C+D+L | 96.75 |
| **State-of-the-art methods** | | |
| Marin et al. 2014 [3] | | 91.28 |
| Marin et al. 2015 [4] | | 96.50 |

techniques can be roughly divided into two main groups: 1) End-to-end fusion, where the representation and the classifier are learned in parallel; and 2) Multi-step fusion, where the representation is first learned and then the classifier is learned from it.

- **End-to-end fusion (*feat_end2end*):** The underlying idea of this approach is to learn an end-to-end deep neural network. In our scenario, the neural network has multiple input-specific pipes, in which each input type is processed by its specific neural net. While colour and depth are both processed by a CNN, the Leap Motion data is processed by a classical neural net with one hidden layer. Then, the last hidden layers of each pipe are concatenated followed by one additional fully connected layer. All the layers are trained together end-to-end (Figure 1 - left).
- **Multi-step (*feat_mstep*):** As in the end-to-end approach, a shared (multimodal) representation vector is created, by concatenating the last hidden layers of each model previous trained individually. Then, for sign recognition, the multimodal representation vector is fed into an additional classifier (Figure 1 - middle).

### 2.2.2 Decision-level fusion

The purpose of decision-level fusion is to learn a specific classifier for each modality and, then, to find a decision rule between them. In this paper, we apply this concept making use of the output class probabilities of the models designed individually for each modality under analysis (see Figure 1 - right). Then, two main kinds of decision rules, to combine these class probabilities, were implemented:

- **Pre-defined decision rules:** Herein, two different pre-defined decision rules were implemented. In the first approach, refereed as *late_mean*, the final prediction is given by the argument that maximizes the averaged class probabilities. In the second approach, refereed as *late_conf*, the final prediction is given by the model with the maximum confidence. The confidence of a model in making a prediction is measured by its highest class probability.
- **Learned decision rule (*late_learn*):** The underlying idea of this approach is to learn a decision rule from the data. Therefore, a descriptor that concatenates the class probabilities, extracted from the individual models of each modality, is created and, then, used as input into a multiclass SVM classifier for sign recognition.

## 3 Experimental Results

The experimental evaluation of the proposed methodologies was performed in a public Microsoft Kinect and Leap Motion hand gesture recognition database with a total of 1400 hand gestures [3, 4]. The dataset was divided into a training set of 1000 images and a test set of 400 images, with signer independence.

The experimental results of the proposed single-modality and multimodal sign recognition methodologies are reported in Table 1 in terms of classification accuracy (Acc). A first observation, regarding single-modality approaches, is that both colour and depth outperform Leap Motion, with accuracies of 93.50%, 91.75% and 82.00%, respectively. However, it should be noticed that Leap Motion sign recognition does not require any kind of preprocessing in order to segment the hand from the background for feature extraction. The most interesting observation is that multimodal fusion often promotes an overall improvement in the sign recognition accuracy. These results clearly demonstrate the complementarity between the three modalities and the potential to tackle the SLR problem via multi-modality. Typically, the classification accuracy increases as each modality is added to the recognition scheme. In particular, the decision-level fusion scheme, with the average decision rule, provides the best overall classification accuracy ($Acc = 97.00\%$). These results outperformed both state-of-art methods [3, 4], with an Acc of 97.00% against 91.28% and 96.50%, respectively.

## 4 Conclusions

This paper addresses the topic of static SLR, by exploring multimodal learning techniques, making use of data from 3 distinct modalities: (i) colour; (ii) depth, both from Kinect; and (iii) Leap Motion data.

Experimental results demonstrate that, in general, multimodal learning techniques outperform single-modality methods. In particular, the proposed decision-level fusion scheme, with the average decision rule, achieved the best results ($Acc = 97.00\%$) and outperforms the current state-of-the-art methods. As future work, it is expected to extend the proposed approaches for dynamic signs.

**Acknowledgements** This work was funded by the Protect "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE010145-FEDER000016 "financed by the North Portugal Regional Operational Programme (NORTE 2020), under PORTUGAL 2020 Partnership Agreement, and through the European Regional Development FUND (ERDF), and also by Fundação para a Ciência e a Tecnologia (FCT) within PhD and BPD grants with numbers SFRH/BD/102177/2014 and SFRH/BPD/101439/2014.

## References

[1] V. Adithya, P. R. Vinod, and U. Gopalakrishnan. Artificial neural network based method for indian sign language recognition. In *Information Communication Technologies, 2013 IEEE Conference on*, pages 1080–1085, 2013.

[2] Fabio Dominio, Mauro Donadeo, and Pietro Zanuttigh. Combining multiple depth-based descriptors for hand gesture recognition. *Pattern Recognition Letters*, 50:101 – 111, 2014. Depth Image Analysis.

[3] G. Marin, F. Dominio, and P. Zanuttigh. Hand gesture recognition with leap motion and kinect devices. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1565–1569, Oct 2014.

[4] Giulio Marin, Fabio Dominio, and Pietro Zanuttigh. Hand gesture recognition with jointly calibrated leap motion and depth sensor. *Multimedia Tools and Applications*, pages 1–25, 2015.

# Automatic Recognition of Students Numbers in Examination Classroom Maps

Aline Peres
aline.peres@tecnico.ulisboa.pt
Joao Rogerio Caldas Pinto
joao.c.pinto@tecnico.ulisboa.pt
Susana Margarida da Silva Vieira
susana.vieira@tecnico.ulisboa.pt

IDMEC
Instituto Superior Tecnico
Universidade de Lisboa
Lisbon, PT

## Abstract

In this work, an algorithm for automatic recognition of handwritten numerals is implemented, applied to examination classroom maps. The algorithm is composed by four main steps: pre-processing, segmentation, feature extraction and classification. For each of these steps, an approach is proposed. In classification, two different algorithms (SVM and CNN) are evaluated in order to combine them into a more robust classification model and improve accuracy. This work also implements some strategies applied in the segmentation process that are based on pre-existing algorithms and can be adapted to others HCR problems. The results presented are from the algorithm applied on real databases (the MNIST and ISTDB, a proprietary database).

## 1 Introduction

The Handwritten Character Recognition (HCR) is a field of research of Optical Character Recognition (OCR), developed in the scope of artificial intelligence, computer vision and pattern recognition. The aim of this work is to find a solution to perform the recognition of students numbers handwritten in an examination classroom map accurately (fig.1 (a)). This requires an accurate segmentation and classification as well as a good preprocessing and feature extraction. A lot of efforts in the field of HCR has been made along the years, existing a lot of well succeed approaches related on literature. However, this is still a problem with many possibilities of improvement and interesting applications.

Several numeral segmentation methods have been proposed contemplating segmentation-recognition and recognition based algorithms. A popular approach was proposed by Fujisawa et al. [2], based on the length of connected components (CC) and its contour information.

In what concerns the image features to be given as an input to some classification algorithm, the Histogram of Oriented Gradients (HOG) descriptors has proved to be suited to the HCR problem. In a comparative study by Surinta et al. [6], the HOG descriptors has shown to perform better for the MNIST database when compared to other well-known techniques like PCA and DCT applied on SVM classifiers.

The Support Vector Machines (SVM) classifiers, developed based on the statistical learning theory of Vapnik [7], has been considered one of the most powerful classifiers for character and numeral recognition [1]. Still, recent accomplishes have been made with Convolutional Neural Networks (CNN) for image classification. It has become one of the most appealing classification methods and has been the key to solve a lot of the most challenging machine learning problems. Moreover, one of the first convolutional network (LeNet 1) was trained on the MNIST database [4]. Also, some novel hybrid CNN-SVM approaches have been designed integrating the synergy of these two classifiers [5].

### 1.1 The database

The datasets used in this work was the MNIST dataset and the ISTDB dataset. The MNIST dataset [3] is a subset of the NIST dataset. It is composed by 60.0000 handwritten training images and 10.000 handwritten test images in gray levels normalized to a 28x28 pixels image.

The ISTDB is a proprietary dataset composed by 3.415 images extracted from 12 classroom maps handwritten by 11 different persons. These images were obtained through the segmentation process related in this work, and had suffer some image processing before feature extraction and classification in order to be as similar as possible to the MNIST dataset samples (fig.1 (b) and (c)).

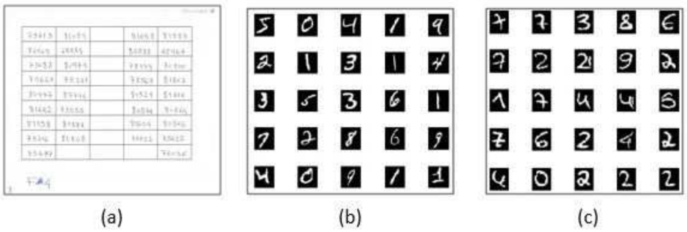


Figure 1: Examples of the classroom maps and of the datasets: (a) Examination classroom map. (b) MNIST. (c) ISTDB.

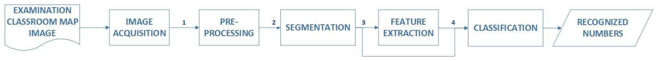


Figure 2: The main algorithm flowchart

## 2 The algorithm

The solution comprises the following main modules: image acquisition, pre-processing, segmentation, feature extraction and classification (fig.2). Notice that the feature extraction module is skipped when feeding a CNN classifier. The input of these classifiers is the raw image.

### 2.1 Pre-processing

The aim of this pre-processing is to identify the regions of interest (ROI) in the input image. In a classroom map, the ROI are all the regions containing a student number. In these terms, the ROIs are all the cells of the map grid so what is expected from the algorithm is to be able to identify the grid and consequently its cells. The algorithm implemented for this task starts by detecting the lines that make up the map grid, despising everything else, and then proceeds to the ROIs identification.

### 2.2 Segmentation

The best way to get individual images of all the individual numerals to be classified is to "cut" the whole number in a cell and then perform its segmentation. So, this task required a sub-module (number segmentation) within the segmentation module, that actually performs segmentation into five individual numerals that make up a student number.

The segmentation algorithm starts by performing a bounding box analysis (BBA) to identify the ROIs (or the cells) and verifies if each of these cells contains a number. If the cell contains a number, then it proceeds to the number segmentation sub-module where a new BBA is performed to the cell area. At this stage, the algorithm verifies if there are any broken, overlapping or touching numerals by accessing the number of objects resulting of the BBA. If there are broken numerals (> 5 objects) the algorithm identifies the disconnected part accessing the object area and merge it to the previous object regarding that arabic numerals are written from the left to the right. If there are overlapping or touching numerals (< 5 objects), the algorithm identifies these objects as well as the number of connected numerals, by computing the area ratio or the length ratio respectively and splits it.

### 2.3 Feature extraction: HOG descriptors

For the computation of the HOG descriptor, the cell size and the block size were set to 8x8 and 16x16 respectively, i.e. each block contains 2x2 cells. Once that the numeral images are 28x28, there are 4 possible positions for

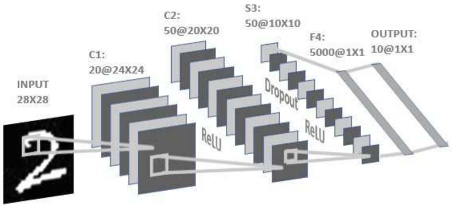


Figure 3: The CNN architecture

the 16x16 blocks and so we get a 36x4=144 dimensional feature vector. It seems like some of the margin pixels of the image are neglected in order to fit the cells to the image size, that can be an issue. The alternative to not neglect any pixel would be to set the cell size to 4x4 and the block to 8x8. This would lead to a 1296 dimensional feature vector that is a much bigger number than the 28x28=784 pixels of the image, violating the dimensionality reduction purpose of the descriptor. Moreover, the margin pixels are almost only background (black pixels) once that the numeral image is centred into a 28x28 box.

### 2.4 Classification: SVM, CNN and combined models

The first classifiers implemented were the CNN in figure 3 fed by the raw image, and a quadratic SVM (polynomial kernel of second order) with *onevsone* coding for the binary learners and fed by the images HOG descriptor vectors. In a attempt to build a powerful classification model, the SVM and a CNN classifiers were combined using two approaches: the same SVM classifier fed by the features extracted from the fully connected layer of the CNN structure and a SVM classifier fed by both the HOG descriptor vectors and the features extracted from the fully connected layer of the CNN structure.

#### 2.4.1 Training and test set

The training set is composed by the 60.000 training samples of the MNIST dataset and other 2.610 samples of the ISTDB corresponding to 8 out of the 12 available classroom maps. The other 805 samples extracted from the remaining 4 maps were used as a test set.

## 3 Results and conclusions

### 3.1 Segmentation

The segmentation algorithm described above were applied to the 12 classroom maps. Each image resulting from the segmentation process was verified manually. The successful rate was of 100% in 9 out of the 12 maps, with two others above 95% and one above 92%. The bad segmented numerals didn't proceed to classification. The overall successful rate was of 99,1%.

It was noticed that the algorithm fails when a number contains a broken numeral but there is no intersection of bounding boxes and in addition the smallest area object corresponds to a single numeral and not to a disconnected part of other object being the reason to some of the errors in segmentation. Moreover, the worst results in the segmentation process were achieved in the maps that the writing was sloppy, i.e., numerals that differs greatly in size in a same number, very tight numbers with a lot of overlapping and touching points between numerals as well as a lot of interruptions in the writing trace leading to broken numerals situations.

### 3.2 Classification

| Classifier | CNN | HOG+SVM | CNN+SVM | HOG+CNN+SVM |
|---|---|---|---|---|
| Acc (%) | 94,04 | 93,42 | 93,17 | 94,66 |

Table 1: Overall performance of the proposed classifiers.

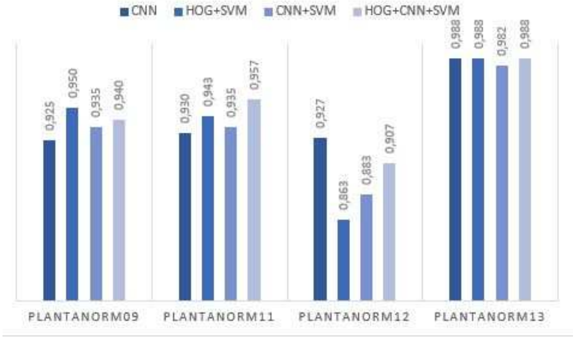


Figure 4: The classifiers performances per maps.

After giving the test set to the classification models in section 2.4 some interesting results were obtained. In table 1 is visible that the overall performance of the HOG+CNN+SVM classifier is slightly better than the other three classifiers. Looking at figure 4 it is notable that although the HOG+CNN+SVM classifier has the best overall performance, for individual maps this is not always true. By comparing the results with its respective maps it was clear that the CNN classifier is more stable, performing accurately even for sloppily written numbers. By other hand, the HOG+SVM classifier can be more accurate for careful maps. The CNN+SVM classifier is clearly a mid term of these two classifiers, revealing more stability than the individual SVM classifier and being more accurate than the individual CNN classifier for some cases. However, it is not so accurate as the HOG+CNN+SVM classifier and so the HOG descriptor vectors should not be overlooked.

## 4 Future work

Considering the performance of the classifiers proposed, an interesting approach would be to design a decision system that would allow the algorithm to take advantage of the strengths of each classifier, by other words a robust ensemble model. Moreover the whole algorithm can be easily adapted to other applications such as the recognition of the student number in an examination sheet. This could be a possible application that would allow the students to have an easier access to their examination sheets avoiding unnecessary dislocations for any revisions.

## 5 Acknowledgements

This work was supported by FCT, through IDMEC, under LAETA -
UID/EMS/50022/2013, IDMEC and Program Investigator
FCT (IF/00833/2014), co-funded by the ESF through the POPH.

## References

[1] Dennis Decoste and Bernhard Schölkopf. Training invariant support vector machines. *Machine learning*, 46(1):161–190, 2002.

[2] Hiromichi Fujisawa, Yasuaki Nakano, and Kiyomichi Kurino. Segmentation methods for character recognition: from segmentation to document structure analysis. *Proceedings of the IEEE*, 80(7):1079–1092, 1992.

[3] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010. URL `http://yann.lecun.com/exdb/mnist/`.

[4] Yann LeCun, LD Jackel, Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, UA Muller, Eduard Sackinger, Patrice Simard, et al. Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 261:276, 1995.

[5] Xiao-Xiao Niu and Ching Y Suen. A novel hybrid cnn–svm classifier for recognizing handwritten digits. *Pattern Recognition*, 45(4):1318–1325, 2012.

[6] Olarik Surinta, Mahir F Karaaba, Tusar K Mishra, Lambert RB Schomaker, and Marco A Wiering. Recognizing handwritten characters with local descriptors and bags of visual words. In *Engineering Applications of Neural Networks*, pages 255–264. Springer, 2015.

[7] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

# Shadow Type Identification for Gait Recognition using Shadows

Tanmay Tulsidas Verlekar[1]
https://www.it.pt/Members/Index/18817

Luís Ducla Soares[2]
https://www.it.pt/Members/Index/511

Paulo Lobato Correia[1]
https://www.it.pt/Members/Index/390

[1]Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Portugal

[2]Instituto de Telecomunicações Instituto Universitário de Lisboa, (ISCTE-IUL), Portugal

## Abstract

Using features acquired from the shadow cast by a walking person can be an alternative for gait recognition whenever the person's body is occluded, such as when capturing images from an overhead position. However, the shadow, depending on the light source characteristics, can be cast as a blob with no distinguishing characteristics. Most state-of-the-art methods fail in the presence of such "diffused" shadows. Thus, this paper presents a novel method to identify the type of shadow cast by a person. The proposed method generates a histogram of the intensity ratio between foreground and background areas, whose analysis allows identifying the type of shadow cast by the person. The proposed method is very promising, achieving a 90% correct shadow type identification with the dataset tested.

## 1 Introduction

Gait is a biometric trait that describes the way a person walks. It is a cyclic combination of movements that results in human locomotion [1]. Thus, in a surveillance environment, gait becomes difficult to hide, unlike other biometric traits such as face, iris or fingerprints. It can also be acquired from a distance, without any cooperation from the person being observed. Such advantages lead to gait recognition attracting significant interest of the research community [1].

### 1.1 State-of-the-art

Traditionally, gait recognition in a surveillance environment, i.e. under the observation of a single 2D camera [2], is performed using appearance based methods such as gait energy image (GEI), gait entropy image (GEnI) or motion energy image (MEI) [3]. These methods rely on spatiotemporal information obtained from the input images to perform recognition – see Fig. 1. Most of these methods acquire gait features from the person's body silhouettes. However, work presented in [4], suggests that gait features can also be acquired from the shadow cast by a walking person.

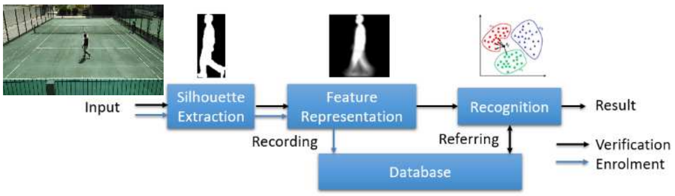


Figure 1: Appearance based gait recognition system.

Several methods that perform gait recognition using shadow silhouettes are reported in the literature, relying on features such as: spherical harmonic coefficients [5], gait contour [6], shadow gait energy image [7], or affine moment invariant coefficients [8]. Among them, methods such as [7] and [8] perform well under surveillance environments being robust to view and appearance changes. The robustness to view change in [7] is achieved by transforming the shadow silhouettes into a canonical view, where the transformation is obtained from low-rank optimization of a gait texture image consisting of shadow silhouettes. Robustness to appearance changes in [8] is achieved by assigning different weights to the altered parts of the shadow GEI.

### 1.2 Motivation

Among the methods presented in the state-of-the-art, it is expected that the shadow cast by a person is always sharp and appears similar to the person's body silhouette. However, this is not always true. In nature, depending on the direction of the rays emanating from the light source, one of the two different types of shadows can be cast. When the light rays travel in the same, well-defined direction, i.e. when the light source is a collimated source of light, the shadow cast by the person is sharp and displays features similar to those that can be acquired from the person's body silhouette – see Fig.2 (a). However, if the light rays have many different directions, multiple overlapping shadow contributions result in a diffused shadow. Such shadow does not possess any distinguishing characteristics of a person – see Fig.2 (b). Also, shadow segmentation methods such as [5] and [7] rely on the feet position of the person to separate the shadow from the person. Since the diffused shadow surrounds the feet of the person– see Fig.2 (b), detection of the feet position becomes impossible. Thus, identifying the type of shadows can help improve performance of the gait recognition systems by allowing them to employ a different shadow segmentation method such as the mixture of gaussians (MOG) [9] in the diffused case.

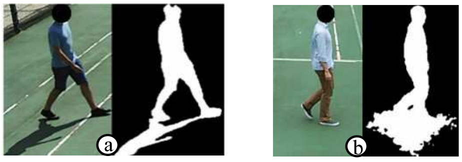


Figure 2: Shadow cast by the person: under a collimated source of light (a) and under a diffused source of light (b).

This paper presents a novel method to identify the type of shadow cast by a person. The method uses the intensity ratio between the foreground and the background to generate a histogram of intensity values. Analysing the plot allows the method to identify the type of shadow cast by the person. Use of the proposed method allows the state-of-the-art gait recognition systems that rely on shadow silhouettes to perform better in surveillance environments.

## 2 Proposed Shadow Type Identification Method

As shown in Fig.1, given an input sequence, the silhouette extraction module relies on background subtraction methods, such as MOG [9] or robust principal component analysis (RPCA) [10] to obtain a foreground mask consisting of the walking person's body and shadow silhouettes. Along with the foreground mask – see Fig.3 (a), these methods also provide a foreground image (FGI) and a mean background image (MBGI), as illustrated in Fig.3 (b) and (c), where the mean background image is estimated from the available video sequence.

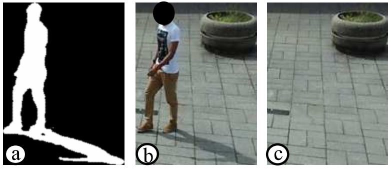


Figure 3: Outputs of background subtraction: foreground mask (a), foreground image (b) and mean background image (c).

To identify the type of shadow cast by a person only the intensity components ($IC$) of $FGI$ and $MBGI$ are used to obtain an intensity ratio $\mu$ for every pixel value $i$, according to (1).

$$\mu_i = FGI_i^{IC}/MBGI_i^{IC} \tag{1}$$

The method then filters the intensity ratio values $\mu$ using the foreground mask to select only those intensity ratio values that either belong to the person's body or to the shadow. A histogram is then generated for the selected pixel values. The histogram plot, depending on the type of shadow cast by the person, can contain one of the two distinct shapes – see Fig.4 (a), Fig.5 (a). The difference in the plot shape is caused by the distribution of intensity ratio values in the sharp and the diffused shadow areas. The person's body area, usually having a significant contrast with the background, results in low intensity ratio values. On the other hand, a sharp shadow contains a significant contrast with its

This work was partially supported by Instituto de Telecomunicações under Fundação para a Ciência e Tecnologia Grant UID/EEA/50008/2013.

background and, being a darker area, also changes the background chromaticity resulting in a large number of low intensity ratio values in the histogram plot – as illustrated in Fig.4.

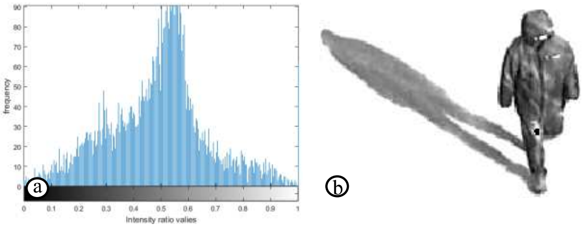


Figure 4: Sharp shadow's intensity ratio: histogram plot (a), for a sample image (b).

In the case of a diffused shadow, a significant contrast still exists between the person's body and the background. However, the diffused shadow only slightly reduces the intensity of the background, not causing significant changes to the background chromaticity. Thus, the diffused shadow generates a large number of high intensity ratio values. Therefore, the resulting histogram plot contains a significant amount of both low and high intensity ratio values, as illustrated in Fig.5.

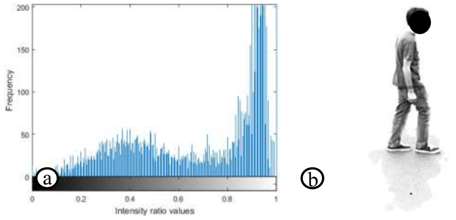


Figure 5: Diffused shadow's intensity ratio: histogram plot (a), for a sample image (b).

To automatically identify the type of shadow, the histogram is organized into 20 bins, followed by the application of a gaussian filter. The bins corresponding to peaks in the histogram are then identified, as illustrated in Fig.6. It is observed that the sharp shadows, i.e. those with low intensity ratio values, are represented by histograms with a single peak – see Fig.6 (a), while diffused shadows, i.e. containing large numbers of both low and high intensity ratio values, are represented by two clearly separated peaks – see Fig.6 (b). Therefore, depending on the number of peaks identified it is possible to have a classification of the type of shadow cast by the walking person.

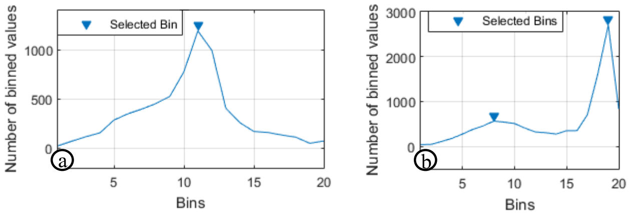


Figure 6: Bin selection for: sharp shadow (a), diffused shadow (b).

## 3 Results

To test the performance of the proposed method, a database is constructed consisting of six people captured across two days. On the first day, the people are recorded under a clear sky and, thus, the shadows cast by the people are sharp. On the second day, people are recorded under a cloudy sky, casting diffused shadows. The recordings are conducted at Instituto Superior Técnico, Lisbon, Portugal, on a tennis court, between 10:00 to 12:00 o'clock in the month of June. Although the recorded people are only six, around 300 frames are captured, as each person walks the full length of the tennis court– see Fig.1 (input). Thus, approximately 3600 frames are available to test the proposed method.

To test the proposed method, all the frames acquired on the first day are indexed as sharp shadow and all the frames acquired on the second day are indexed as diffused shadow. Next, the proposed method is applied to individual frames and the resulting output is matched against the index to obtain the correct identification rate.

Using the proposed method, a correct shadow type identification of 90% is achieved. During the tests, the proposed method is applied to individual frames, however, gait features are usually acquired from an integer number of complete gait cycles, each with a typical duration between 1 and 2 seconds. Thus, using a voting policy the shadow type can be identified for the entire duration of the gait cycle, further improving the proposed method's results.

In the results obtained, it is observed that some of the errors result from some sharp shadows being represented by histograms with two peaks. However, in most of these error cases, the distance between the two peaks is extremely small, allowing easy improvement in the results.

Another source of error happens when the person's attire blends in with the background, with a camouflage effect, due to the lack of contrast, resulting in high intensity ratio values in the person's body region. The proposed method is unable to distinguish between the two types of shadows under such conditions.

## 4 Conclusion

A person can casts either a sharp or a diffused shadow depending on the light source characteristics. A sharp shadow is cast under a collimated source of light, which is characterised by the shadow's appearance being similar to a person's body silhouette. A diffused shadow is cast when the light rays from a light source travel in different directions. In this case, the shadow appears as blob around the person with no distinctive characteristics. Most state-of-the-art gait recognition systems that rely on shadows cannot operate with diffused shadows. Similarly, some state-of-the-art shadow segmentation methods cannot separate diffused shadows from the person's body. Under such conditions, identifying the type of shadow can be useful.

The paper proposes a novel method to identify the type of shadow. The method generates a histogram of ratios computed between the foreground and background intensity values. The two types of shadows are represented with two distinct plots. Sharp shadow is saturated with low intensity ratio values, while diffused shadow contains a significant amount of both low and high intensity ratio values. Thus, the shadow type can be identified by analysing these histograms.

A contrast between the person's body and the background allows the proposed method to perform extremely well. However, when the person camouflages with the background the proposed method performs poorly. To tackle, this the future work includes the use of textures and other colour channels that better distinguish the person, shadow and the background. To further test the proposed method, the database should be expanded to include more people and diverse conditions.

## References

[1] Y. Makihara, D. S. Matovski, M. S. Nixon, J. N. Carter and Y. Yagi. Gait recognition: Databases, representations, and applications. *Wiley Encyclopedia of Electrical and Electronics Engineering*. 2015.

[2] T. Verlekar, P. Correia and L. Soares, View-Invariant Gait Recognition System Using a Gait Energy Image Decomposition Method. *IET Biometrics*, 6(4): 299 - 306, 2017,

[3] Z. Lv, X. Xing, K. Wang and D. Guan. Class energy image analysis for video sensor-based gait recognition: A review. *Sensors*, 15(1): 932-964, 2015.

[4] A. Stoica. Towards Recognition of Humans and their Behaviors from Space and Airborne Platforms: Extracting the Information in the Dynamics of Human Shadows. In *Proc. BLISS*, 2008.

[5] Y. Iwashita, A. Stoica, and R. Kurazume. Person Identification using Shadow Analysis. In *Proc. BMVC*, 2010.

[6] Y. Iwashita, A. Stoica, and R. Kurazume. Finding people by their shadows: Aerial surveillance using body biometrics extracted from ground video. In *Proc. EST*, 2012.

[7] T. Verlekar, P. Correia and L. Soares. Gait Recognition Using Normalized Shadows. In *Proc. EUSIPCO*, 2017.

[8] Y. Iwashita, A. Stoica, and R. Kurazume. Gait Identification Using Invisible Shadows: Robustness to Appearance Changes. In *Proc. EST*, 2014.

[9] P. KaewTraKulPong and R. Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. *Video-based surveillance systems* 1: 135-144, 2002

[10] P. Rodriguez and B. Wohlberg. Fast principal component pursuit via alternating minimization. In *Proc ICIP*, 2013.

# Classification of High Resolution Aerial Images Using Deep Learning

Dmytro V. Solovey[1]
dmytro.solovey@tecnico.ulisboa.pt
João C. Pinto[2]
joao.c.pinto@tecnico.ulisboa.pt
Alexandra Moutinho[2]
alexandra.moutinho@tecnico.ulisboa.pt
Pedro Gamboa[3]
pgamboa@ubi.pt

[1] Academia Militar
Rua Gomes Freire
1169-203 Lisboa, Portugal

[2] IDMEC, Instituto Superior Técnico
Av. Rovisco Pais
1049-001 Lisboa, Portugal

[3] AeroG, Universidade da Beira Interior
Calçada Fonte do Lameiro
6201-001 Covilhã, Portugal

## Abstract

This work has as main objective of developing a method and algorithm that can extract and classify the types of terrain, the existence of constructions and objects of interest in high resolution aerial photography. It is a possible alternative for drawing fuel maps automatically, a very important step in the context of fire prevention, the goal of the project "FIRECAMP2" within which this work was developed. The proposed methodology is composed by the following steps: segmentation, image processing and classification. For the segmentation the superpixels algorithm (SLIC) was used. After segmentation the images needs to be processed to isolate the superpixels. A pre-trained deep neural network was adapted for this purpose. A training set was created, with five classes, that allowed to train the adapted neural network. Finally, is performed the classification of the superpixels. The results obtained in the classification of five aerial images are between 77% and 88% of accuracy. The confusion matrices show that in certain classes it was possible to obtain precision values of 96%. The confusion occurs between two very similar classes when both classifications can be considered correct. The merging of these two classes could allow global precision values above 90%.

## 1 Introduction

With the emergence of Drones and "UAV" equipped with the latest technologies, it became possible to obtain high-resolution aerial images. This increases the need to processing the images automatically. In scope of the "FIRECAMP2" project, Almeida [2] presents a methodology to model the fire propagation, in which is necessary to create fuel maps. The methodology developed in this work can be used for this purpose. In this work the intention is to segment and classify aerial images in five classes. For this a training set was created and a pre-trained deep neural network was adapted and trained. The automatic semantic segmentation used in this work allows to increase the speed of classification and interpretation of aerial images and, at the same time avoiding the subjectivity of the manual process.

## 2 Creating Training Dataset

For this work it was necessary to create a database containing labeled images corresponding to each of the five proposed classes. Several aerial images were segmented with SLIC algorithm [1, 4] into superpixels. Using image processing techniques it was possible to isolate each of them. The superpixels that best represented each class were chosen and labeled manually. In total the dataset receive 1074 manually labeled images for training, with the following distribution by classes:

| Class | Quantity |
|---|---|
| Water | 202 |
| Building | 76 |
| Road | 150 |
| Vegetation | 521 |
| Land | 125 |

Table 1: Number of images by classes.

It is important not only the number os examples in the dataset, but also its distribution. In this case, it was possible to extract 521 labeled images for vegetation class but, only 76 for buildings. In the consequence its expected that the vegetation class will be classified better.

## 3 Classification Method

For classification was used a deep convolutional artificial neural network AlexNet [5, 6, 7]. This network consists of 25 layers. Convolutional, pooling, ReLU, dropout, softmax and fully conected layes is a part of this network. It was pre-trained to distinguish 1000 different classes. In this work the network was adapted to this specific problem. The layer 23 is replaced with one fully connected with only 5 neurons. Each neuron represents our classes. The learning rate parameter was increased to the last layers and other hyper-parameters were defined [3]. The maximum value obtained for the mini batch size (learning parameter) was 64. This parameter depends of a processing capacity. The network was re-trained, with the training set previously prepared, to distinguish only five proposed classes. To make this possible its also necessary to replace the last layer of the CNN.

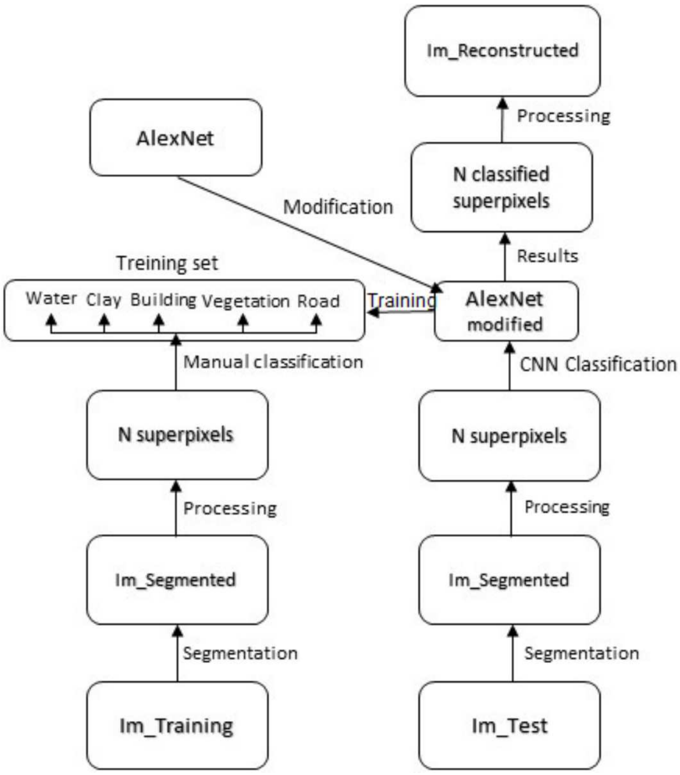


Figure 1: Metodology graph.

## 4 Metrics

The technique to measure the accuracy was based on a direct comparison between the colors of the pixels that results directly from the classification by the algorithm with the respective ground truth image. The global accuracy value is obtained from (eq.1).

$$Accuracy = \frac{N^o\ of\ correct\ classified\ pixels}{Total\ n^o\ of\ pixels} \qquad (1)$$

Confusion matrices are also presented to analyze the accuracy obtained by each class.

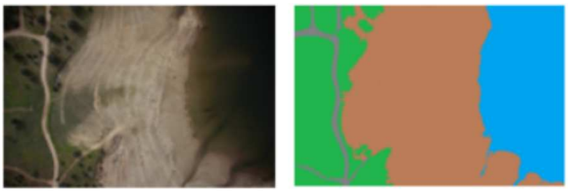


Figure 2: Aerial image with corresponding ground truth.

## 5 Results

In the segmentation process the SLIC algorithm was used for the generation of superpixels. The aerial images provided were taken from approximately the same altitude. This made possible to adjust the parameter that sets the amount of super pixels to 400. Its important to choose the right value for this parameter to increase the ability to extract features. In the Figure 2. We can see a part of aerial image after segmentation with good separation between different classes.

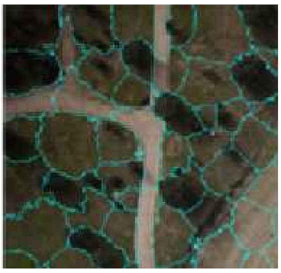


Figure 3: Part of aerial image after segmentation.

The CNN was trained using GPU. This has substantially reduced the time required for training. In this work the training took 2235s, limited by the available processing capacity. Also the depth of CNN was not fully exploit. However, the obtained results are higher than 95% in some classes. A high error in the classification is due to the similarity between two classes. A merge between these classes can increase overall accuracy to values over 90%. After the classification the superpixels are again merged to form a classified image (Figure 3).

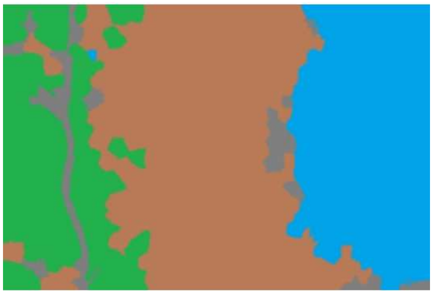


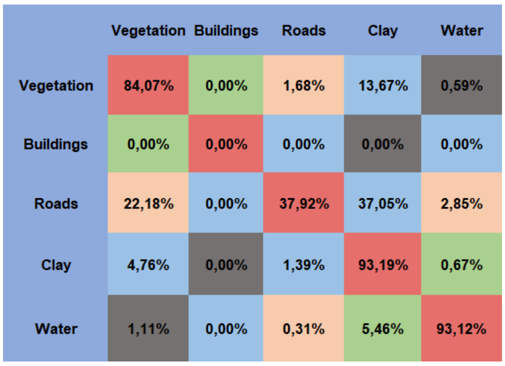
Figure 4: The classified image obtained by the algorithm.

The confusion matrix (Figure 4) obtained for the image in Figure 1 is very similar to the rest of the obtained results. Accuracy above 93% for classes such as land and water. The vegetation also has a high value of 84%. The difficulty in elaborating ground truth is another factor that influences the results. The lowest obtained value is for class of roads. If we consider that in the case of dirt roads the classification is correct in both cases, the accuracy will increase to 74.97%. The overall accuracy obtained was 87.4% which is close to the best values presented in the state of the art.

In Table 2 are represented the global accuracy obtained for five aerial images which were chosen for tests. The lowest result is 77.27% and in the same time, at the same image was obtained 96.76% of accuracy in buildings extraction.

| Image | Global Accuracy |
|---|---|
| 1 | 87.4% |
| 2 | 88% |
| 3 | 77.27% |
| 4 | 83.82% |
| 5 | 79.97% |

Table 2: Results for all processed images.



| | Vegetation | Buildings | Roads | Clay | Water |
|---|---|---|---|---|---|
| Vegetation | 84,07% | 0,00% | 1,68% | 13,67% | 0,59% |
| Buildings | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% |
| Roads | 22,18% | 0,00% | 37,92% | 37,05% | 2,85% |
| Clay | 4,76% | 0,00% | 1,39% | 93,19% | 0,67% |
| Water | 1,11% | 0,00% | 0,31% | 5,46% | 93,12% |

Figure 5: The confusion matrix obtained for the Figure 1.

## 6 Conclusions

This paper presents a methodology for classifying high resolution aerial images through deep learning. Several steps are required for the process, among them: creation of database for training, segmentation, adaptation of a pre-trained CNN and classification. The results obtained show that in some classes the obtained accuracy is above 90% and exists a large percentage of confusion between the land and road classes. Future work will deal with a deeper study of the steps of this methodology. Increase the training set, increase the depth of learning, use new techniques for segmentation.

## 7 Acknowledgments

This work was supported by FCT, through IDMEC, under LAETA - UID /EMS/50022/2013.

## References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchina, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 34(11):2274–2282, November 2012. doi:10.1109/TPAMI.2012.120.

[2] M. Almeida, L. Ribeiro, D. Viegas, J. Azinheira, A. Moutinho, J. Pinto, J. Barata, K. Bousson, J. Silva, M. Martins, R. Ervilha, and J. Pereira. Analysis of fire hazard in campsite areas. *Fire Technology*, March 2016. doi:10.1007/s10694-016-0591-5.

[3] J. Bergado. A deep feature learning approach to urban scene classification. Master's thesis, University of Twente, March 2016. www.itc.nl/library/papers2016$/msc/gfm/bergado.pdf$.

[4] O. Linares. Segmentação de imagens de alta dimensão por meio de algoritmos de detecção de comunidades de super pixels. Master's thesis, USP, June 2013. http://www.teses.usp.br/teses/disponiveis/55/55134/ tde-25062013-100901/publico/OscarQuadrosrevisada.pdf.

[5] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *Computer Vision and Pattern Recognition (CVPR)*, October 2015. doi:10.1109/CVPR.2015.7298965.

[6] L. Shao, Z. Cai, L. Liu, and K. Lu. Performance evaluation of deep feature learning for rgb-d image/video classication. *Information Sciences*, January 2017. doi:10.1016/j.ins.2017.01.013.

[7] S. Srinivas, R. Sarvadevabhatla, K. Mopuri, and N. Prabhu. Deep learning for medical image analysis. pages 25–52, 2017. doi: 10.1016/B978-0-12-810408-8.00003-1.

# Segmentation of citizen ID cards for identity verification in access control to public events

Marina Castro, Cristiana Carpinteiro, Joana Rodrigues, Margarida Fernandes, Ana Costa, Diogo Goncalves
Faculty of Engineering of the University of Porto (FEUP)

Joao C. Monteiro
INESC TEC

## Abstract

Biometric recognition has become a growing trend in recent years due to the inherent limitations of more traditional identification techniques. Access to events by comparison of the ID card with its bearer has become a common approach and its automation would represent an interesting improvement. In the present paper we propose a new system to automatically extract face and text data from the ID card to allow this automation. Some interesting results were observed in segmentation of multiple regions of interest, while building opportunity for a wide diversity of future work.

## 1 Introduction

In modern society, identity verification is required in numerous day-to-day circumstances, ranging from security checks in airports and banks, to workplace admittance and ticket selling and validation. The main problem with conventional methods is that they don't prevent individuals from impersonating others, since they rely, almost exclusively, on possession of identity documents [4]. The solution should therefore involve a non-transferable method of identification, such as biometric authentication [2, 5]. This solution should be of practical use and low cost, so that it can be used in several environments with diverse conditions.

Biometric authentication is already in place in some events and tourist attractions. It is used to match individuals to photos in a preexisting database, recognizing possible security threats [1] or providing efficient ticket verification at events and preventing individuals from impersonating others [4]. The system we propose can be used for the latter purpose. The main difference proposed in this work is that, while the existing mechanism requires a preexisting photo of the individual, ours does not, with both data sources for the verification process being acquired in real-time during its use.

Our proposed system captures real-time images from video footage, capturing the user's face as well as extracting data from his/her identity card. By segmenting the document, through recognition of the chip, face (Viola-Jones algorithm), and adjusting its shape and orientation, we can obtain information from the ID card and validate the identity of the holder of said document. A face comparison algorithm, not yet implemented, can then ascertain if the subject matches the identity of the presented ID card, eliminating the transferability issue of these documents. Furthermore, optical character recognition (OCR) methodologies are used, in parallel to the face verification module, to allow for a second layer of identity verification through the analysis of the textual data present in the machine-readable zone (MRZ) of the ID card, as well as similar data extracted from the frontal part.

## 2 Proposed Methodology

### 2.1 System Overview

The functional pipeline of the proposed system is depicted in Figure 1. To the present moment most of the presented work has been carried out on the ID card segmentation (including a layout-based detection of multiple regions-of-interest (ROIs)) and OCR methodology for the MRZ region.

### 2.2 ID Card Segmentation

The segmentation of the ID card was accomplished using a color-based segmentation in the *b* channel from CIE L*a*b color space, since the cards have a distinctive color. The CIE L*a*b color space was chosen for its perceptually oriented color representation [3]. The Otsu method was used for the thresholding of this segmentation. Morphologic opening was performed in order to smooth the image and identify the main object due to its markedly quadrilateral shape. The binary image resulting from this first segmentation was used to correct the perspective of the ID photo in order to have a perfect shaped rectangle.For this perspective correction, a 2-D spatial transformation, using a transformation matrix defined by 4 control points, was performed. The selection of the four control points was carried out automatically, using the corners of the binary image. The corners were detected based on the boundary pixels location. To check whether the presented face of the card is the front or back one, the presence of a face (photography) and a chip and their spatial relation are the acquired parameters.

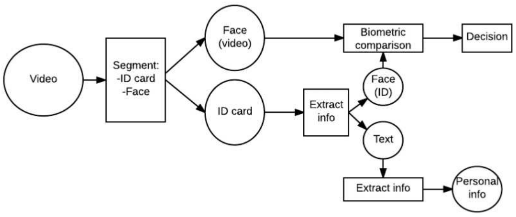


Figure 1: Functional pipeline of the proposed system.

### 2.3 ROIs Detection

To extract the textual information from both the frontal and back side of the ID card, a region-based segmentation was considered. Face detection is performed in the frontal side, so that the subject's photo can be compared with the acquired video stream, while the position of the face can be used as an anchor to the spatial localization of the remaining ROIs, such as text boxes. In order to detect the position of the face in the card. the Viola-Jones algorithm was chosen. This detection algorithm uses a cascade classifier to decide whether a window that slides over the image contains the object of interest. The chip was also chosen as a second anchor for ROI detection, due to its distinctive coloration, that allowed for a simple detection via k-means clustering in the *a* and *b* channels of the aforementioned L*a*b color space. After face and chip detection, the text boxes in the ID card, such as the ones with the subject's name and parenthood, were detected by following the standard layout of the Portuguese ID card, as depicted in Figure 2. Regarding the back of the card we simply considered the respective lower part of the card for the OCR algorithms.

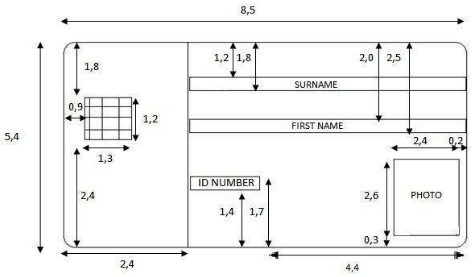


Figure 2: Layout of the front side of the Portuguese ID citizen card.

### 2.4 Optical Character Recognition

Having the card correctly detected, the MRZ is segmented knowing its typical position in the back of the Portuguese ID. Then, a bottom-hat operation (difference between the result of a closing operation and the original image) is applied to the region, taking as structuring element a disk whose ray is the hundredth part of the length of the diagonal of the image. This creates a method of enhancing the characters that is adapted to each card image, considering its scale.The bottom-hat operation allows the enhancement of the contrast of regions which are darker than their surroundings, such as the characters in the MRZ.

For the recognition of MRZ characters, an algorithm provided by MATLAB was used, which returns an object that contains recognized text, text location, and a metric indicating the confidence of the recognition result. To improve the results of the detection, it was specified that the characters were contained in a single block of text and the possible set of characters used in the MRZ was defined as the only possible set of recognition options.

## 3 Results and Discussion

### 3.1 Dataset

The developed algorithms were tested on a small dataset of ID card images acquired from 17 volunteers. Acquisition conditions were kept variable between individuals so as to better replicate real-life scenarios. Some examples of such images will be presented throughout this sections as the results of each developed block are outlined. [1]

### 3.2 Results

Some of the obtained results are depicted in Figures 3 to 6, representing, respectively, examples of the ID card segmentation, ROIs detection and OCR text extraction. As it can be observed, the developed algorithms present some interesting results in this preliminary dataset. Even though no integration of the diverse blocks was implemented, the individual performance of each one represents a promising starting point for future work. It has to be noted that the whole ROIs and OCR blocks are limited by the correction of the card orientation to a perfectly horizontal position. This step seems to be working correctly but might need further exploration for more severe orientation variations.

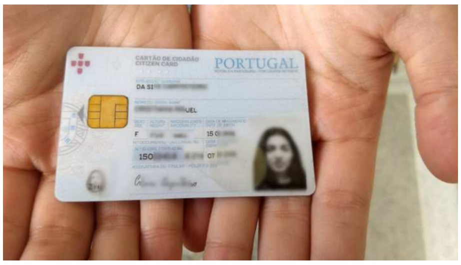


Figure 3: Example of Portuguese ID card from the developed dataset.

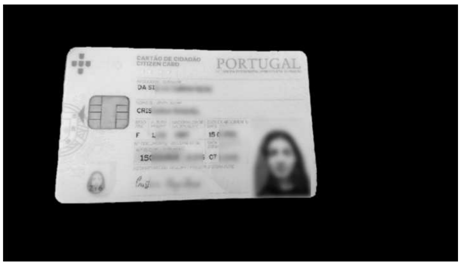


Figure 4: Segmentation result on the L channel of the L*a*b colorspace.

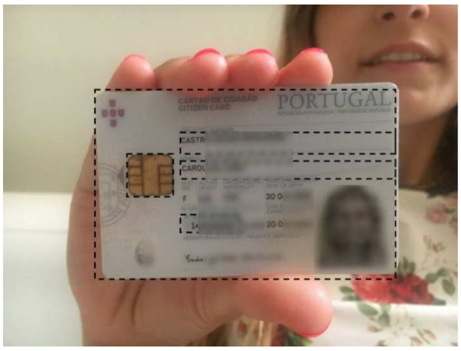


Figure 5: ROIs detection example (face, chip and text boxes).

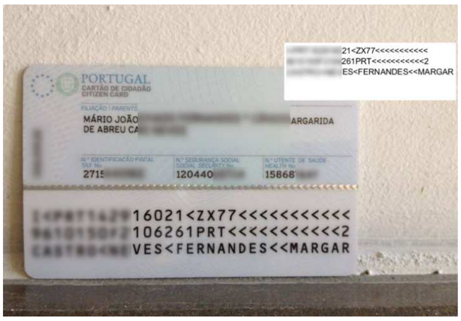


Figure 6: Example of OCR result on the MRZ region.

## 4 Conclusion and Future Work

In the present paper, some preliminary work on an identity verification system for event access was outlined. The main goal of the system in its final shape is to allow the identity of subject to be verified both via face verification between video stream and the photo in his/her ID card, as well as by an additional check via the information in the MRZ region of the ID card.

The developed modules of ID card segmentation and face/text information extraction presented promising results but a multiplicity of improvements can still be implemented. First, no integration of the multiple modules is developed. There is some preliminary work in the development of a graphical interface, but we are still on the early stages of its development. Furthermore, the face recognition block is still inexistent and is, obviously, needed if any real-world application is to be developed in the future. All these improvements will be the focus of work in the immediate future.

## References

[1] Find Biometrics. Nec facial recognition helps secure sky-high tourist attraction, 2017.

[2] M. Mizoguchi H. Imaoka and M. Hara. Biometrics technology to preserve safety and security. *Journal of Information Processing*, 51-12:1547–1554, 2010.

[3] Ingmar Lissner and Philipp Urban. Toward a unified color space for perception-based image processing. *IEEE Transactions on Image Processing*, 21(3):1153–1168, 2012.

[4] Akitoshi Okumura, Takamichi Hoshino, Susumu Handa, Yugo Nishiyama, and Masahiro Tabuchi. Identity verification of ticket holders at large-scale events using face recognition. *Journal of Information Processing*, 25:448–458, 2017.

[5] Y. Seto. Trends and prospects in biometric security authentication technology. *Journal of Information Processing*, 47-6:571–576, 2006.

[1] All individuals signed a written consent for their data to be used for scientific research purposed only.

# Segmentation based approaches for face detection on thermal images

Ricardo F. Ribeiro
rfribeiro@ua.pt
António J. R. Neves
an@ua.pt

DETI/IEETA, University of Aveiro,
3810-193 Aveiro, Portugal

## Abstract

Infrared cameras or thermal imaging cameras are devices that use infrared radiation to capture an image. This work proposes the development or adaptation of several methods for face detection on infrared thermal images. The well known algorithm developed by Paul Viola and Michael Jones, using Haar feature-based cascade classifiers, is used to compare the traditional algorithms developed for visible light images when applied to thermal imaging. In this paper, we present three different methods for face detection. In the first one, an edge detection algorithm is applied to the binary image and the face detection is based on these contours. In the second method, a template matching method is used for searching and finding the location of a template image with the shape of human head in the binary image. In the last one, a matching algorithm is used. This algorithm correlates a template with the distance transform of the edge image. We performed some tests using the different algorithms. The results show that the proposed methods have promising outcome, but the Template Matching algorithm is the most suitable for the performed experiments.

## 1 Introduction

In the electromagnetic spectrum, the visible light spectrum is the only part that the human eye can see. Due to the fact that infrared radiation is invisible to the human eye, thermal cameras use infrared sensors to capture that radiation, transforming it into visible images.

The use of thermal infrared cameras has been increasing in various scientific areas. A survey providing an overview of the current applications is presented in [3].

The infrared thermal camera used is a FLIR LEPTON long-wavelength infrared (LWIR) (50° shutterless) camera module with a focal plane array of $80 \times 60$ active pixels. This camera is connected to a Raspberry Pi 3 to be used as processing device. The image acquisition process is based on a project developed by the company Pure Engineering [2]. Figure 1(a) shows an example of a thermal image.

In this work, we propose algorithms for face detection using thermal infrared cameras.

This paper is organized as follows. In Section II, we present the techniques that have been proposed for face detection. In Section III, we provide experimental results. Finally, in Section IV, we draw some conclusions.

## 2 Proposed Approach

In this section, different developed algorithms and methods for face detection on thermal images are presented.

### 2.1 Haar Cascades

Haar Cascades is a machine learning approach, using Viola and Jones algorithm [9], for visual object detection where a cascade function is trained from positive images (images of faces) and negative images (images without faces). The performance of the trained classifier will be better, as more images are used.

### 2.2 Implementation Details

Thermal image is, on a first stage, segmented and filtered with morphological operators in order to obtain a binary image for the later use of the proposed algorithms. Segmentation uses the Otsus method that is a thresholding binarization method [8] and filtering is performed using morphological operators, such as dilation, erosion, opening and closing [4]. Examples are shown in Figure 1(b) and (c), respectively.

In this work, the following algorithms are developed/adapted and implemented:

- **Face Contours** - Acquisition and filtering of the contours in order to obtain the longest contour in the binary image and detect the face through it.
- **Template Matching** - Technique for finding areas of an image that match to a template image [5].
- **Chamfer Matching** - Technique to find the best alignment between two edge maps [6].

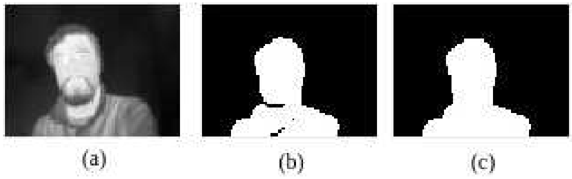


Figure 1: In (b) the segmentation of the thermal image (a). The result of applying the morphological operation is presented in (c).

## 3 Experimental Results

In this section, the experimental results to verify the effectiveness of the implemented algorithms are present.

### 3.1 Haar Cascades

It was made a Haar Cascade training for face detection in thermal images. The dataset used, in this attempt of training classifiers for face detection on thermal images, is OTCBVS Benchmark Dataset Collection [1].

Examples of the face detection using the trained classifier are presented in Figure 2. In 1602 faces in the images, 3829 possible faces were detected, but only 1464 corresponded to true positive. The precision and recall are 38.1% and 91.4%, respectively.

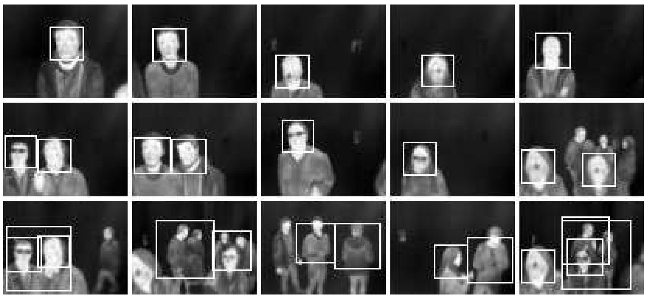


Figure 2: Examples of the detection using Haar Cascade algorithm with the improved classifier for thermal images.

### 3.2 Proposed Methods for Face Detection

In this section, experimental results are presented using a pre-processing of the thermal images for the three proposed algorithms.

#### 3.2.1 Face Contours

This algorithm was developed for single face detection. Figure 3 shows the sequence of images processed by this algorithm, from image capture to final face detection. These results can be influenced if the contours of the face are discontinued for some reason.

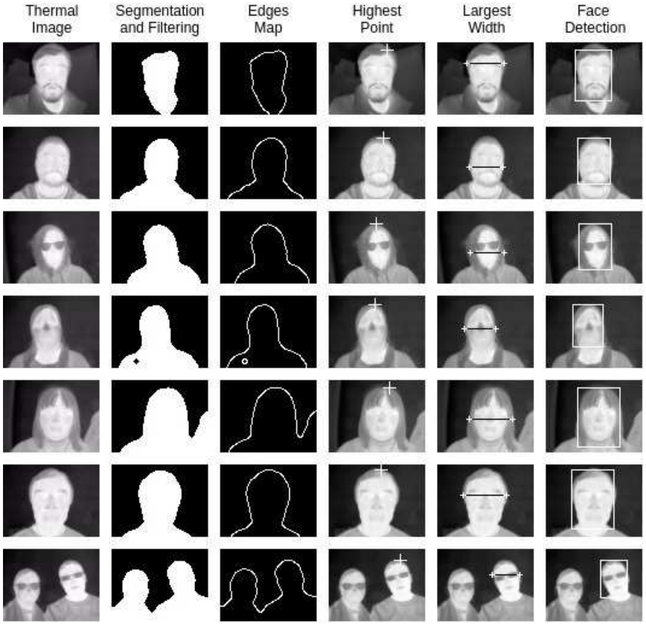


Figure 3: Examples of the results using Face Contours algorithm.

### 3.2.2 Template Matching

When using template matching algorithm a good template is needed. Figure 4 shows the used template in different scales.

Figure 5 shows some examples of the face detection using this algorithm. In 3229 faces in the images, 3535 possible faces were detected, but only 2770 corresponded to true positive. The precision and recall are 78.4% and 85.8%, respectively.

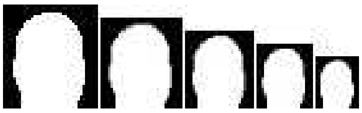


Figure 4: Example of the template in different scales.

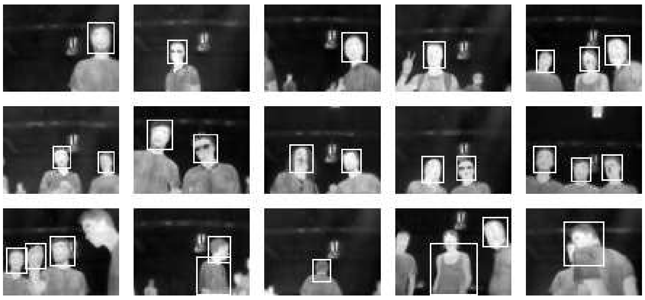


Figure 5: Face detection based on Template Matching.

### 3.2.3 Chamfer Matching

This algorithm uses the same template of the Template Matching algorithm, shown in Figure 4, and the same captured thermal images to perform a test. Some results applying the Chamfer Matching are shown in Figure 6.

In 3229 faces in the images, 3361 possible faces were detected, but only 1810 corresponded to true positive. The precision and recall are 53.9% and 56.1%, respectively.

## 3.3 Processing Time

Table 1 shows the average of the processing time of the tests for each algorithm.

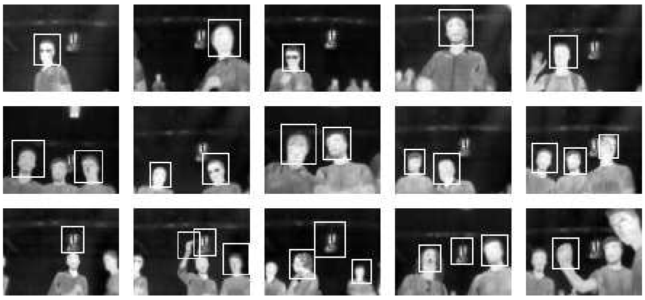


Figure 6: Examples of the face detection using Chamfer Matching algorithm.

| Proposed Algorithms | Average Processing Time (ms) |
|---|---|
| Haar Cascade | 87.92 |
| Face Contours | 90.11 |
| Template Matching | 90.41 |
| Chamfer Matching | 242.93 |

Table 1: Processing time of each proposed algorithm.

## 4 Conclusion

Haar Cascade using Viola and Jones algorithm has better performance and accuracy based on [7]. Face detection in thermal image using Haar Cascade can be improved using one of the proposed methods for image segmentation and create several thermal images for training database in order to obtain better detection.

Comparing the algorithms used in this work, Template Matching is the most suitable. Although Face contours and Haar Cascade have a processing time similar to Template Matching, the Haar Cascade has a lower precision, detecting few faces and Face Contours has the disadvantage of not detecting multiple faces. Chamfer Matching has a similar detection to Template Matching for single detection but the processing time of this algorithm is almost 3 times slower and the detection performance, when applied to multiple face detection, decreases significantly.

## References

[1] Otcbvs benchmark dataset collection. `http://vcipl-okstate.org/pbvs/bench/`.

[2] Pure Engineering. Flir lepton breakout board. `http://www.pureengineering.com/projects/lepton`. Retrieved April 2017.

[3] Rikke Gade and Thomas B Moeslund. Thermal cameras and applications: A survey. *Machine vision and applications*, 25(1):245–262, 2014.

[4] Henk JAM Heijmans. Connected morphological operators for binary images. *Computer Vision and Image Understanding*, 73(1):99–120, 1999.

[5] John P Lewis. Fast normalized cross-correlation. In *Vision interface*, volume 10, pages 120–123, 1995.

[6] Ming-Yu Liu, Oncel Tuzel, Ashok Veeraraghavan, and Rama Chellappa. Fast directional chamfer matching. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1696–1703. IEEE, 2010.

[7] Jiři Mekyska, Virginia Espinosa-Duró, and Marcos Faundez-Zanuy. Face segmentation: A comparison between visible and thermal images. In *Security Technology (ICCST), 2010 IEEE International Carnahan Conference on*, pages 185–189. IEEE, 2010.

[8] Reza Farrahi Moghaddam and Mohamed Cheriet. Adotsu: An adaptive and parameterless generalization of otsu's method for document image binarization. *Pattern Recognition*, 45(6):2419–2431, 2012.

[9] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

# On the Suitability of Light Field Imaging for Road Surface Crack Detection

David Fernandes, Paulo Lobato Correia
Instituto Superior Técnico, Universidade de Lisboa /
Instituto de Telecomunicações

Henrique Oliveira
Instituto Politécnico de Beja / Instituto de Telecomunicações

## Abstract

During traditional road surveys, inspectors capture images of pavement surface using cameras that produce 2D images, which can then be automatically processed to get a road surface condition assessment. In this paper the use of a light field imaging sensor is proposed, notably the Lytro Illum camera, to explore whether the richer information captured by this imaging sensor provides additional cues useful to improve the automatic detection of road surface cracks. The preliminary results obtained indicate the interest in further exploring the disparity information captured by the light field sensor.

## 1 Introduction

This paper explores the potential of using of a light field camera, the Lytro Illum, to improve the automatic detection of road surface cracks, when compared to the usage of conventional 2D cameras.

Light field cameras are emerging as powerful sensor devices that capture the full spatio-angular visual information in a viewing range. This means that the light field information can be processed to obtain a matrix of 2D images, the sub-aperture images, corresponding to slightly different perspectives of the scene. This allows exploring the disparity between the various sub-aperture images, potentially making road cracks more salient and easier to detect.

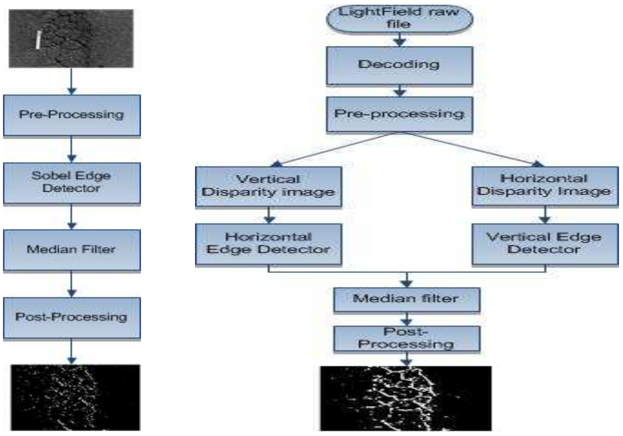
The matrix of sub-aperture images can be computed from the raw light field with the help of the Matlab Light Field Toolbox [1].

To evaluate the potential of light field imaging for crack detection, a small dataset was captured in conditions similar to those considered during traditional road pavement surveys, with the camera positioned at 1 *m* from the pavement surface, with its optical axis perpendicular to the road pavement.

This paper considers two simple crack detection systems: (i) a simple 2D crack detection system, *2D_Crack_Detector*, which is used for comparison purposes; and (ii) the proposed light field crack detector *LF_Crack_Detector*, which uses the matrix of sub-aperture images computed form the light field information. Both systems use the same image processing techniques to obtain the final crack detection results, without focusing on optimizing the performance of the systems, but rather focusing on the initial detection stage, to evaluate the potential of exploring the light field disparity information for improving the detection of cracks.

## 2 Crack Detection

This section details the main modules composing each of the crack detection systems considered. Section 2.1 details the conventional 2D crack detector pipeline, whose architecture is shown in Figure 1 (a), while Section 2.2 details the proposed light field system – see Figure 1 (b).



(a) (b)

*Figure 1: System architecture (a) 2D_Crack_Detector, (b) LF_Crack_Detector*

### 2.1 2D Crack Detector

The considered 2D system follows the architecture of Figure 1(a). A simple set of techniques were considered for illustration purposes, which can later be improved, for instance considering those presented in [2][3] . The system takes as input a grayscale image, and then follows the steps briefly explained here, and whose results are illustrated in Figure 2: **(1) Saturation,** since it is assumed that crack pixels are darker than non-crack pixels, those pixels with high intensities can replaced with an intensity value that is clearly above the crack intensities, but allowing to reduce the intensity variance. **(2) Sobel edge detector**, a Sobel mask is applied to detect the edges in the image, although it often also amplifies the existing noise. **(3) Median Filter**, with a window size of 5x5 pixels, is used to remove some of the image noise. **(4) Post-Processing**, considers a Gaussian Blur filter, with a standard deviation value of 2, to further reduce noise and to soften the detected crack edges. Finally, a thresholding operation is applied to generate the image where candidate crack pixels are identified. The threshold value used is 25, all the pixels with intensity value above 25 are identified as belonging to a crack.

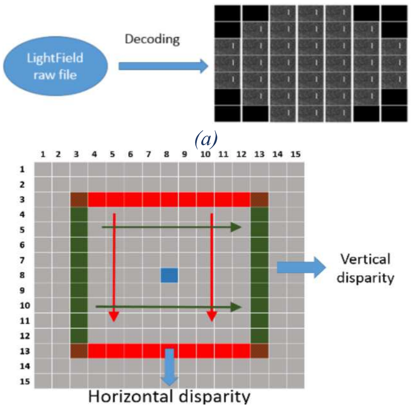
*(a) (b) (c) (d) (e)*

*Figure 2: Results of each processing step of the 2D system: (a) Original Image, (b) Step 1, (c) Step 2, (d) Step 3, (e) Step 4. (Images are shown with increased brightness for easier visualization)*

### 2.2 Proposed Light Field Crack Detector

Since the raw light field (LF) includes more information than a simple 2D image, the proposed LF crack detector exploits the captured disparity for improving the detection of cracks.

The **Decoding** step, takes the raw light field and creates a 15x15 matrix of 2D sub-aperture images, each with spatial resolution 435x625 pixels and representing a slightly different perspective – see Figure 3 (a).



(a)

(b)

*Figure 3: (a) Decoding step and the sub-aperture matrix; (b) Step (3) - Vertical and Horizontal disparity computation illustration.*

The **Pre-processing** step has same purpose as the saturation step in section 2.1, i.e., to remove some of the image noise.

To exploit the disparity present in the light field images, a selection of sub-aperture images is considered after some experimentation, and also according to [4] considering the difference between images at position symmetrical to the central sub-aperture image. Figure 3 (b) illustrates the **Vertical (resp. Horizontal) Disparity Image** creation, when subtracting the images of column (resp. row) 3 with those of column (resp. row) 13 of the matrix, and summing the obtained differences.

Figure 4 shows the summation results when considering different amounts of disparity, i.e. columns (resp. rows) closer or further apart from each other. The number of the used sub-aperture images considered in each column (resp. row) can also be varied, as illustrated in Figure 6.

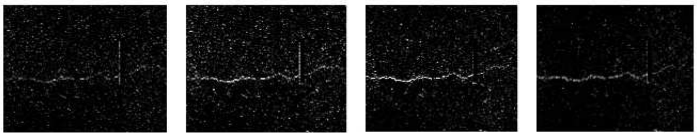
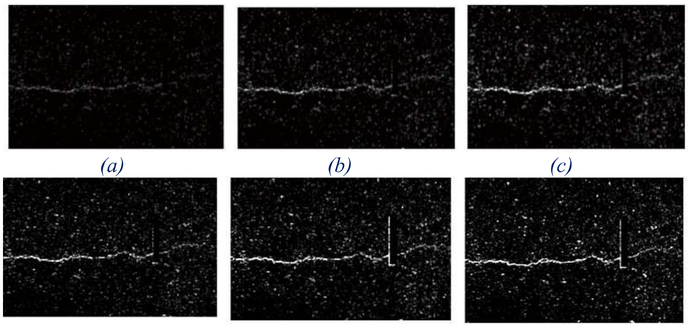

(a) (b) (c) (d)

Figure 4: (a) $7^a$ and $9^a$ rows and columns, (b) $5^a$ and $11^a$ row and columns (c) $3^a$ and $13^a$ rows and columns (d) $1^a$ and $15^a$ rows and columns


(a) (b) (c)

(c) (d) (e)

*Figure 6: Number of Images used (a) 1 images, (b) 3 images, (c) 5 images, (d) 7 images, (e) 9 images, (f) 11 images.*

Summation of the horizontal and vertical disparities is illustrated in Figure 7 (left and right), capturing the crack information present in the specified direction.

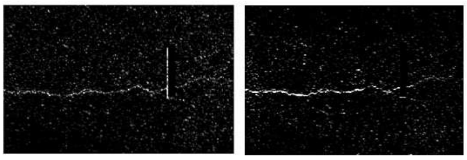


*Figure 7: Sum of the horizontal differences (left), sum of the vertical differences (right).*

Eventually, an **edge detector** can be applied to the horizontal and vertical disparity images to use the same processing pipeline as applied to the 2D images, thus enhancing the crack details. The sum of both images - see Figure 8 (right) is the best capturing all the crack information present in the image.

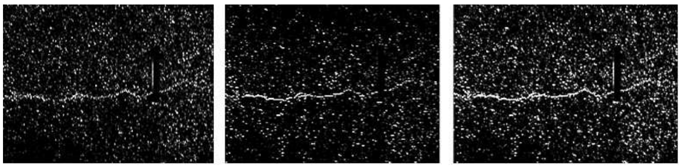


*Figure 8: Step 3: Vertical mask (left), Horizontal mask (centre), Sum of the previous images (right).*

The **edge detector** enhances the crack details but it also increases the noise present in the non-crack areas.
To reduce the effect of noise, the same non-linear filtering technique, **2-D median filter**, with a window size of 5x5 pixels, can be applied, as illustrated in Figure 9.

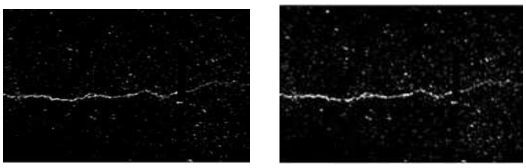

(a) (b)

Figure 9: Median filter applied: (a) without the Sobel step; (b) after the Sobel detector.

The final step, **Post-Processing**, performs a thresholding operation with a chosen threshold value of 64, to identify the crack pixels, after applying a Gaussian filter, with a standard deviation value of 2, to smooth the noise in the non-crack areas. The results are illustrated in Figure 10.

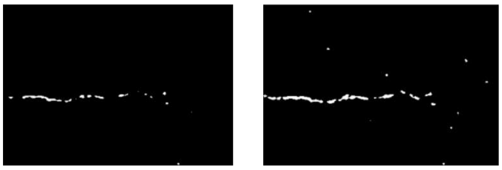

(a) (b)

Figure 10: Results of the post-processing step (a) without and (b) with the application of the Sobel detector.

These results show that a transversal crack can be easily identified in Figure 10. The threshold values and other parameters used are the same during the processing of both the images. The difference between them is due to the usage of an edge detector, adopted to enhance the edges, although the presence of groups of pixels presenting small dimension in the image may occur.

## 3 Discussion and Conclusions

Figure 11 includes a set of additional results considering both the 2D and the proposed light field crack detection systems.

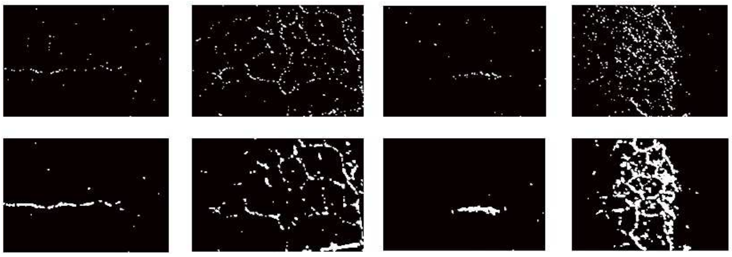

(a) (b) (c) (d)

*Figure 11: Results of both systems on 4 images (first row: 2D crack detector; second row: light field crack detector): (a) Image 1, (b) Image 2, (c) Image 3, (d) Image 4*

As a conclusion, the usage of light field imaging, with some basic image processing techniques, seems to provide better crack detection results, than using the same techniques over a 2D image (in this case the central 2D sub-aperture image). Exploring the light field disparity information showed a better definition of the cracks present in the test images.
Future work involves considering a more sophisticated and complex image processing and classification techniques in the proposed system architecture. Also, the acquisition of a larger test image dataset and an elaboration of a quantitative measure like the F-measure metric will be addressed in a further work.

## References

[1] D. G. Dansereau, "Light Field Toolbox for Matlab," vol. 34, no. 2, pp. 2013–2015, 2013.

[2] H. Oliveira and P. L. Correia, "Automatic road crack detection and characterization," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 155–168, 2013.

[3] R. Amhaz, S. Chambon, J. Idier, and V. Baltazart, "Automatic Crack Detection on Two-Dimensional Pavement Images: An Algorithm Based on Minimal Path Selection," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 10, pp. 2718–2729, 2016.

[4] A. Sepas-Moghaddam, P. Correia and F. Pereira, "Light field local binary patterns description for face recognition," in *IEEE ICIP*, Beijing, China, Sep. 2017

## Acknowledgement

This work has been partially supported by Instituto de Telecomunicações under Fundação para a Ciência e Tecnologia, Grant UID/EEA/ 50008/2013.

# Pre-processing approaches to improve Facial Verification in Unconstrained Environments

Daniel P. F. Lopes
lopesdaniel@ua.pt
Antonio J. R. Neves
an@ua.pt

DETI/IEETA, University of Aveiro,
3810-193 Aveiro, Portugal

## Abstract

Face Recognition has been received great attention over the last years, not only on the research community, but also on the commercial side. One of the many uses of face recognition is its use on access control systems where a person has one or several images of his/her face associated to an Identification Document, also known as facial verification. This process however will most likely be performed under an unconstrained environment which makes it difficult to achieve good performance results.

In order to improve the verification performance, this paper presents two pre-processing approaches: a face image selection according to the yaw movement of the head of the subject and a camera calibration algorithm that will adjust the camera gain and exposure in order to ensure that the face image sent to the verification process will have the best quality by not losing any of the major features that defines it.

While comparing the camera calibration method proposed with a calibration method that solely involves the automatic exposure of the camera, verification results improved quite substantially achieving a 15% improvement. As for the image selection according to the head pose, the results had a 40% improvement.

## 1 Introduction

Face verification is one of the two tasks that can be done with face recognition. It is performed when it is taken a photo of an unknown person with a claim of identity, deciding whether the person is who he/she claims to be by comparing the photo taken to the photos database linked to the claim of identity. This process is specially useful in situations when a person is entering a private space and he/she carries some sort of identification with him/her.

Traditionally, this process is done manually, involving a person (i.e. a security guard) doing the verification. However, over the years, there is appearing some interest on the automation of this process. Although there are many advantages, such as faster and less expensive systems, there are still many open problems in order to make this system viable and reliable since these systems may be implemented in unconstrained environments such as an outdoor environment.

One of the ways to improve the reliability of these type of systems is to improve its verification performance. In order to achieve this, there are presented in this paper two pre-processing techniques that will help to improve the quality of the images that serve as inputs of the verification system.

## 2 Environment for the Tests and Algorithms studied

For the test of methods that are further proposed, some tests were done in order to create a dataset of face images. The tests were done on the entrance of a building with a camera that was at a distance of 1.5 meters from the ground. People that were entering the building were asked if they wanted to contribute for the tests. As people agree to join, they would stand at about 1 meter from the camera and then, face images were taken in order to create a dataset to test the pre-processing techniques. The algorithm used for face detection was the Histogram of Oriented Gradients (HoG) [5] provided by *dlib* [2] and for the face recognition it was tested the Local Binary Patterns [3] provided by *OpenCV* [1]. Finally, the camera used for the tests was a *uEye UI-640LE-C-HQ* along with a 12mm lens.

## 3 Calibration

The algorithm proposed is a mixture between the calibration of exposure and gain. First, the exposure time is set to automatic (feature provided by the camera) in order to adapt to the light and the environment conditions. Once the exposure time is set and if a face is detected by the face detector, the gain will be adjusted according to the pixels inside of the face bounding box. This gain adjust is made through a solution presented in [6] which consists in the calculation of the Mean Sample Value (MSV) of the gray values of the image histogram. Equation 1 shows the calculation of the MSV. The gain is increased if the MSV calculated is below 2.4 and decreased if it has a value superior than 2.6.

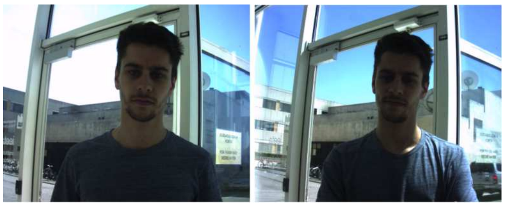


Figure 1: An example of two images acquired for face detection and further construction of the dataset. On the left the image was acquired at 9AM and on the right the image was acquired at 2PM of the same day.

$$MSV = \frac{\sum_{j=0}^{4}((j+1)x_j)}{\sum_{j=0}^{4} x_j}, \quad (1)$$

where $x_j$ is the sum of the gray values in the region of the histogram (in the proposed approach it is divided the histogram into five regions).

Figure 2 shows the comparison between the camera calibration made with only auto-exposure and the method proposed.

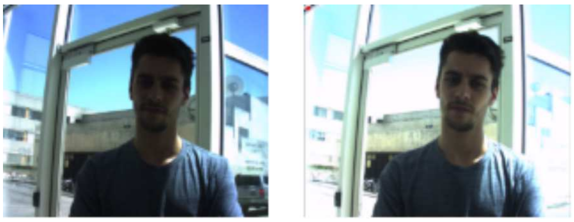


Figure 2: Image acquired using auto-exposure (on the left) and using the calibration method proposed (on the right).

## 4 Head Pose for Image Selection

Head pose at the time that an image is acquired is crucial not only for detection, but also for recognition. There are three degrees of freedom that a human head can move that are shown on Figure 3.

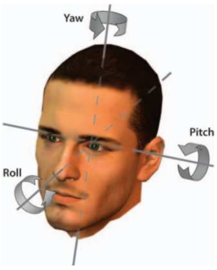


Figure 3: The three degrees of freedom of a human head can be described by the egocentric rotation angles pitch, roll, and yaw [4].

As there is any of these movements, recognition rates start to get lower. In order to calculate these type of movements, the *68 face landmarks* provided by *dlib* are used. The approach proposed focuses on the yaw movement and it assumes symmetry on a face of a person so the distance from the tip of the nose to the left ear is the same from the right ear to the tip of the nose. This, if detected by the camera means that there is no yaw movement and the person is faced to the camera. However, if those two distances differ, that means that the head has some yaw movement towards the camera thus the face image will not be chosen. An example of the head of the subject with and without yaw movement is presented on Figure 4.

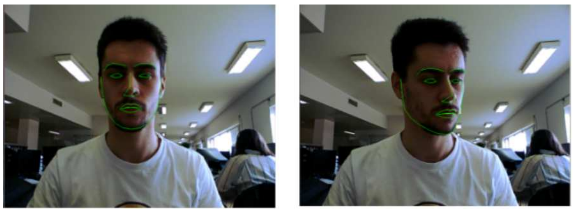


Figure 4: Head of the subject with no yaw movement (image accepted on the verification system) and with yaw movement (image not accepted on the verification system).

## 5 Experimental Results

Once explained the methods proposed and the dataset used to study them, experimental results are presented in order to analyse the behaviour of the recognition algorithms under an unconstrained environment.

There were acquired in total around 1500 face images. 350 of them were used for the construction of the database while the rest of the 1150 were used for the comparison process. The output of these comparisons are "confidence values" which were used to construct the ROC (Receiver Operating Characteristic) curves that are further presented which help to observe the performance of the pre-processing techniques presented.

### 5.1 Camera Calibration proposed

Figure 5 presents the results using LBP algorithm for facial verification using the calibration with just automatic exposure and the method proposed.

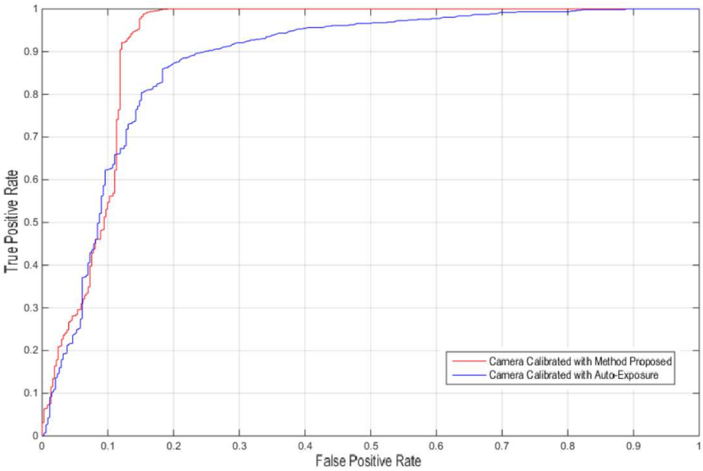


Figure 5: ROC curves showing the results of LBP algorithm performance between the auto-exposure calibration and the calibration proposed.

Observing Figure 5 it is possible to notice a significant improvement with the new calibration algorithm. When using it, it is possible to set a threshold that will have a True Positive Rate (TPR) of 100% for just 20% of False Positive Rate (FPR).

### 5.2 Head Pose for Image Selection

Figure 6 presents some of the face images that are accepted and some that are not using this pre-processing technique.

As for Figure 7, it shows the performance of the LBP recognition algorithm comparing verification results with the image selection method and without it. Noteworthy to mention that these tests were done with the calibration method proposed as it presented better results than the auto-exposure calibration method.

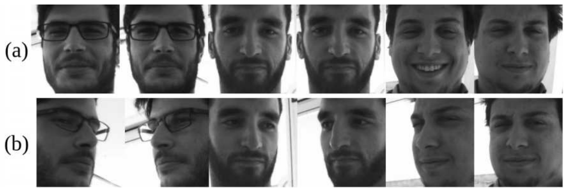


Figure 6: Face images that are accepted in the verification system (a) and face images that are not accepted (b).

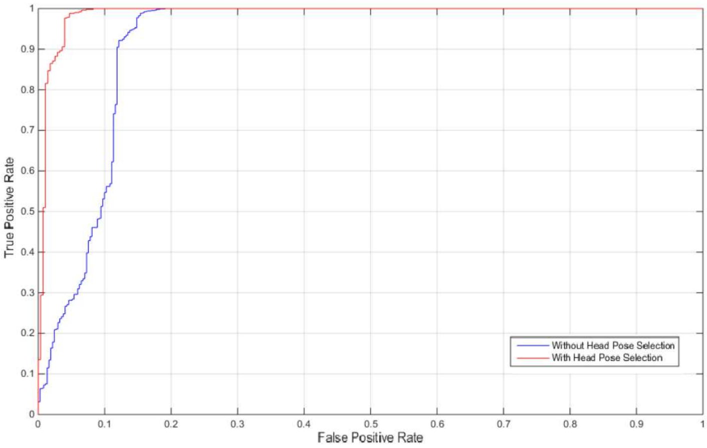


Figure 7: ROC curves showing the results of LBP algorithm performance with and without the image selection method.

Looking at Figure 7, the LBP algorithm suffers a major increase in performance since without the method proposed in order to have a 100% TPR the system would accept 20% of FPR and with the method proposed, it would drop the FPR to just 5%.

## 6 Conclusions

Both of the pre-processing methods presented improve the verification results when compared to the traditional approaches. Although the better results, certain aspects are needed to take into account when using these new methods such as processing time and computational capacity needed to run them.

Nevertheless, the tests made and their results show that there is some potential on the use of these techniques for verification systems that are working in unconstrained environments which have a big demand on the commercial branch.

## References

[1] Opencv. URL `http://opencv.org/`.

[2] dlib. URL `http://http://dlib.net/`.

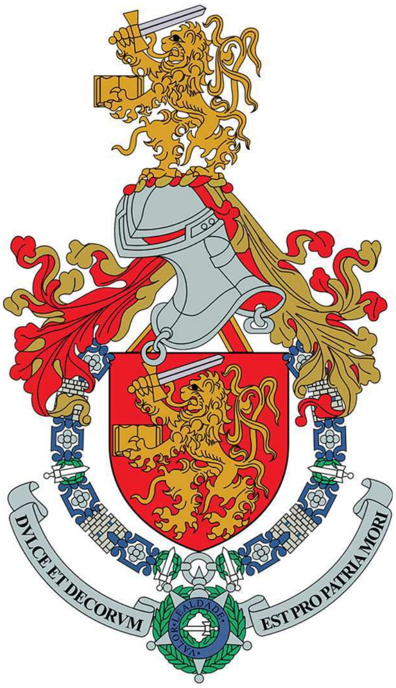
[3] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12): 2037–2041, 2006.

[4] Euclides N Arcoverde Neto, Rafael M Duarte, Rafael M Barreto, João Paulo Magalhães, Carlos Bastos, Tsang Ing Ren, and George DC Cavalcanti. Enhanced real-time head pose estimation system for mobile device. *Integrated Computer-Aided Engineering*, 21 (3):281–293, 2014.

[5] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[6] António JR Neves, Bernardo Cunha, Armando J Pinho, and Ivo Pinheiro. Autonomous configuration of parameters in robotic digital cameras. pages 80–87, 2009.

# Author Index


DVLCE ET DECORVM
EST PRO PATRIA MORI

academia
militar
# um desafio para o futuro

## CURSOS

Os cursos de formação de Oficiais ministrados na Academia Militar conferem o grau de mestre através de um ciclo de estudos integrados e reflectem as necessidades das diferentes Armas e Serviços do Exército e da Guarda Nacional Republicana, nas quais o aluno ingressa após a conclusão do curso.

## MESTRADOS INTEGRADOS

Exército

- Ciências Militares – Infantaria
- Ciências Militares – Artilharia
- Ciências Militares – Cavalaria
- Administração Militar
- Engenharia Militar
- Engenharia Electrotécnica Militar – Transmissões
- Engenharia Electrotécnica Militar – Material
- Engenharia Mecânica Militar
- Medicina
- Medicina Dentária
- Medicina Veterinária
- Ciências Farmacêuticas

Guarda Nacional Republicana

- Ciências Militares – Infantaria
- Ciências Militares – Cavalaria
- Administração
- Engenharia Militar
- Engenharia Electrotécnica Militar – Transmissões
- Engenharia Mecânica Militar
- Medicina
- Medicina Veterinária
- Ciências Farmacêuticas