International Journal of Interactive Multimedia and Artificial Intelligence

September 2022, Vol. VII, Number 6
ISSN: 1989-1660

unir LA UNIVERSIDAD EN INTERNET

"Nobody phrases it this way, but I think that artificial intelligence is almost a humanities discipline. It is really an attempt to understand human intelligence and human cognition."
Sebastian Thrun

Special Issue on New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence

INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE

ISSN: 1989-1660 –VOL. 7, NUMBER 6

EDITORIAL TEAM

Editor's Note

Due to important advances in technologies such as artificial intelligence, big data, the Internet of Things or bioinformatics produced in recent years, it is necessary to conduct a thorough review of current ethical patterns. One of the research fields that is in full expansion and with a broad future is technology ethics or tech ethics. Just a few years ago, this type of research was considered niche, with very few technology researchers involved. At present, due to the explosion of new applications of artificial intelligence, their problems and their legal barriers have helped innumerable initiatives, declarations, principles, guides, and analyses to flourish, focused on measuring the social impact of these systems and on the development of a more ethical technology. This is, therefore, a problem that needs to be addressed from an academic and multidisciplinary point of view, where experts in ethics and behavior work together with experts in new and disruptive technologies.

The international conference "Disruptive Technologies Tech Ethics and Artificial Intelligence" (DITTET) provides a forum to present and discuss the latest scientific and technical advances and their implications in the field of ethics. It also provides a forum for experts to present their latest research in disruptive technologies, promoting knowledge transfer. It provides a unique opportunity to bring together experts in different fields, academics, and professionals to exchange their experience in the development and deployment of disruptive technologies, artificial intelligence, and their ethical problems.

DITTET intends to bring together researchers and developers from industry, humanities, and academia to report on the latest scientific advances and the application of artificial intelligence, as well as its ethical implications in fields as diverse as climate change, politics, economy or security in today's world.

This Special Issue contains extended versions of selected works presented at the 1st International Conference on Disruptive Technologies, Tech Ethics and Artificial Intelligence (DiTTEt 2021), held in Salamanca (Spain) in September 2021 [1].

Juan F. de Paz Santana[1]
Gabriel Villarrubia González[1]

[1] University of Salamanca

References

[1] J. F. de Paz Santana, D. H. de la Iglesia, A. J. López Rivero, Eds., *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence*, Springer Cham, 2022, doi: https://doi.org/10.1007/978-3-030-87687-6.

DOI: 10.9781/ijimai.2022.09.008

TABLE OF CONTENTS

OPEN ACCESS JOURNAL

ISSN: 1989-1660

The International Journal of Interactive Multimedia and Artificial Intelligence is covered in Clarivate Analytics services and products. Specifically, this publication is indexed and abstracted in: *Science Citation Index Expanded, Journal Citation Reports/ Science Edition, Current Contents®/Engineering Computing and Technology.*

A Clustering Algorithm Based on an Ensemble of Dissimilarities: An Application in the Bioinformatics Domain

Manuel Martín Merino*, Alfonso José López Rivero*, Vidal Alonso, Marcelo Vallejo, Antonio Ferreras

Computer Science School, Universidad Pontificia de Salamanca, Salamanca (Spain)

Received 7 May 2022 | Accepted 4 July 2022 | Early Access 19 September 2022

Abstract

Clustering algorithms such as k-means depend heavily on choosing an appropriate distance metric that reflect accurately the object proximities. A wide range of dissimilarities may be defined that often lead to different clustering results. Choosing the best dissimilarity is an ill-posed problem and learning a general distance from the data is a complex task, particularly for high dimensional problems. Therefore, an appealing approach is to learn an ensemble of dissimilarities. In this paper, we have developed a semi-supervised clustering algorithm that learns a linear combination of dissimilarities considering incomplete knowledge in the form of pairwise constraints. The minimization of the loss function is based on a robust and efficient quadratic optimization algorithm. Besides, a regularization term is considered that controls the complexity of the distance metric learned avoiding overfitting. The algorithm has been applied to the identification of tumor samples using the gene expression profiles, where domain experts provide often incomplete knowledge in the form of pairwise constraints. We report that the algorithm proposed outperforms a standard semi-supervised clustering technique available in the literature and clustering results based on a single dissimilarity. The improvement is particularly relevant for applications with high level of noise.

Keywords

Bioinformatics, Clustering, Kernel Methods, Machine Learning, Metric Learning.

DOI: 10.9781/ijimai.2022.09.007

I. Introduction

CLUSTERING algorithms such as k-means depend heavily on finding an appropriate dissimilarity that reflects accurately the object proximities [1]. This depends on the nature of the data and project requirements [2]. In practice, a wide range of dissimilarities may be defined based for instance on different features of the objects [3], [4]. Different dissimilarities lead often to significant changes in clustering results. Some researchers have addressed this problem learning a general distance from the data [5] but this is a challenging task for high dimensional applications [6]. Therefore, instead of considering a single distance metric an appealing approach is to learn a combination of dissimilarities from the data.

Several authors have developed learning algorithms for multiview clustering that are able to integrate a set of dissimilarities obtained from different features of the objects [7]. Following the same approach [8], Hu et al. [9] have proposed multiple kernel k-means clustering algorithms that might consider a set of dissimilarities using the kernel trick. However, these learning algorithms are unsupervised and may not provide metrics that help to increase the cluster separability [6].

* Corresponding author.

E-mail addresses: mmartinmac@upsa.es (M. Martín Merino), ajlopezri@upsa.es (A. López).

For certain Bioinformatics applications, weak supervised information is available in the form of which pairs of proteins or genes are related [10]. This incomplete supervision may be incorporated into semi-supervised clustering algorithms formulated as pair-wise constraints [11], [12]. Must-link constraints when x_i and x_j belong to the same cluster and cannot-link constraints when x_i and x_j belong to different clusters.

Some researchers have proposed algorithms to learn the metric from a set of equivalence constraints based on the Mahalanobis distance [1], [6], [13]. However, they are based on a single metric that may not be appropriate for certain applications and do not perform well with high dimensional data with noise. Besides, they are prone to overfitting and are computationally intensive due to the large number of parameters involved. Other non-linear metric learning approaches have been developed based on kernel methods [14], [15]. Again they are based on a single dissimilarity and suffer from similar drawbacks.

In this paper, we follow the approach of multiple kernel clustering algorithms [16], [17], that learn a combination of kernels to improve the clustering results. However, this kind of researches relies on complex optimization algorithms and often are not designed to incorporate supervised information in the form of pairwise constraints. The main contribution of this paper is to propose a novel semi-supervised clustering algorithm that learns an ensemble of dissimilarities from incomplete knowledge in the form of pairwise constraints. The problem is formulated as learning the combination of

multiple kernels (similarities) that maximizes the separability among the clusters considering the pairwise constraints. The loss function is convex and quadratic without local minima and it is optimized in dual space efficiently. Besides, it incorporates a penalty term to control the complexity of the family of distances avoiding the overfitting.

The algorithm has been evaluated using several benchmark UCI data sets and two problems of cancer samples identification based on the gene expression profiles. The empirical results suggest that the method proposed improves the clustering results obtained considering a single dissimilarity and a standard supervised clustering method proposed by Xing et al. [13] that learns the metrics from pairwise constraints.

This paper is organized as follows: Section II presents the clustering algorithm proposed that learns a combination of dissimilarities using pairwise constraints. Section III illustrates the performance of the algorithm using several benchmark and two complex cancer samples identification datasets. Section IV discusses the contributions of this paper in the context of related work. Finally, Section V gets conclusions and outlines future research trends.

II. Material and Methods

In this section we present the semi-supervised clustering algorithm developed based on an ensemble of dissimilarities and the experimental datasets considered. First, sections A and B introduce the kernel version of k-means clustering algorithm and the empirical kernel map, that allow us to extend a kernel clustering algorithm to work with a given dissimilarity. Thus, the problem of learning a linear combination of dissimilarities may be formulated as learning a linear combination of kernels. Next, in section C an idealized kernel is defined for clustering applications that helps to reduce the intra-cluster distances while increasing the inter-cluster separability considering the available pairwise constraints. Section D presents the learning algorithm for the linear combination of kernels that best approximate the idealized kernel, subject to a set of pairwise constraints. Section E comments the meaning of the non-null Lagrange multipliers in the dual space as support vectors. Finally, section F describes the features of the benchmark and cancer datasets considered.

A. Kernel K-means Clustering

Let $X = \{x_1, x_2, \ldots, x_n\} \in \Re^d$ be the training set, Z_{ki} the clustering indicator matrix defined as 1 if x_i belong to cluster k and 0 otherwise. k-means clustering looks for a set of representatives $\{c_k\}_{k=1}^{C}$ and a partition of the objects into C groups that minimize the sum of square distances to the cluster representatives:

$$\min_{Z\in[0,1]} \sum_{k=1}^{C}\sum_{i=1}^{n} Z_{ki}\|\mathbf{x}_i - \mathbf{c}_k\|^2 \tag{1}$$

$$\text{s. t} \quad \sum_{k=1}^{C} Z_{ki} = 1 \tag{2}$$

This error function is optimized by an iterative algorithm in two steps. First the centroids for each cluster are computed, next each object is assigned to the group corresponding to the nearest centroid according to the euclidean distance. The use of the euclidean distance induces a bias towards spherical groups. K-means clustering has been extended to more general dissimilarities by mapping non-linearly the original samples to a high dimensional reproducing kernel Hilbert space $\mathcal{F}$ [9]. Let Φ be the non-linear mapping to feature space $\mathcal{H}$. Kernel k-means optimizes the following sum of square errors in the reproducing kernel Hilbert space:

$$\min_{Z\in[0,1]} \sum_{k=1}^{C}\sum_{i=1}^{n} Z_{ki}\|\Phi(\mathbf{x}_i) - \mathbf{c}_k\|_{\mathcal{H}}^2 \tag{3}$$

$$\text{s. t} \quad \sum_{k=1}^{C} Z_{ki} = 1 \tag{4}$$

where $c_k = \frac{1}{n_k}\sum Z_{ki}\,\Phi(x_i)$ is the centroid for cluster k in the kernel feature space. Considering that in this feature space $\Phi^T(x_i)\Phi(x_j) = K(x_i, x_j)$, the L_2 norm can be written exclusively in terms of kernels evaluations as:

$$\|\Phi(\mathbf{x}_i) - \mathbf{c}_k\|_{\mathcal{H}}^2 = K(\mathbf{x}_i,\mathbf{x}_i) - \frac{2}{n_k}\sum_{j=1}^{n} Z_{kj}K(\mathbf{x}_i,\mathbf{x}_j) + \frac{1}{n_k^2}\sum_{j=1}^{n}\sum_{l=1}^{n} Z_{kj}Z_{kl}K(\mathbf{x}_j,\mathbf{x}_l) \tag{5}$$

The optimization of the square error function (3) in the feature space can be solved by algorithm 1.

Algorithm 1. Kernel k-means algorithm

1: **Inputs** K: kernel matrix, C: number of clusters

2: ***Initialize***: The C clusters $C_1^{(0)}, \ldots, C_C^{(0)}$

3: Set t = 0

4: For each x_i compute the cluster with the nearest centroid: $k^*(\mathbf{x}_i) = \text{argmin}_K \|\Phi(\mathbf{x}_i) - \mathbf{c}_k\|^2$ using (5)

5: Update the clusters $C_k^{(t+1)} = \{\mathbf{x}_i \mid k^*(\mathbf{x}_i) = k\}$

6: Go to step 3 and update t = t + 1 if not converged

7: **Return**: $C_1, \ldots, C_C$ partitioning of the objects

B. The Empirical Kernel Map

We have mentioned earlier that the learning algorithm for the kernel k-means clustering can be written exclusively in terms of kernel evaluations. For certain applications only a dissimilarity matrix is available and it is often difficult to obtain a vectorial representation for the data. Therefore, the dissimilarity should be incorporated into the algorithm directly through the kernel definition. To this aim, we first map the dissimilarity to a feature space where the dot product defines a Mercer kernel [18]. Depending on the kernel definition, the map may transform linearly or non-linearly the original distance given rise to a wider family of dissimilarities. Next, we introduce the empirical kernel map proposed by [19].

Let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a dissimilarity and $R = \{x_1, x_2, \ldots, x_n\}$ a subset of representatives drawn from the training set. The mapping to embed a given dissimilarity to a feature space is defined as:

$$\begin{aligned} \Phi: \mathcal{X} &\to \mathcal{F} \\ \mathbf{z} &\to \Phi(\mathbf{z}) = [\phi_1(\mathbf{z}), \phi_2(\mathbf{z}), \ldots, \phi_n(\mathbf{z})] \end{aligned} \tag{6}$$

where

$$\Phi(\mathbf{z}) = D(\mathbf{z}, R) = [d(\mathbf{z},\mathbf{x}_1), d(\mathbf{z},\mathbf{x}_2), \ldots, d(\mathbf{z},\mathbf{x}_n)] \tag{7}$$

This mapping Φ embeds the dissimilarity into a functional Hilbert space where feature j is given by $d(\,.\,, x_j)$. The number of representatives considered determines the dimensionality of the feature space. Now, the dot product in feature space defines the kernel for a given dissimilarity:

$$k(\mathbf{z},\mathbf{z}') = \langle\phi(\mathbf{z}),\phi(\mathbf{z}')\rangle = \sum_{j=1}^{n} d(\mathbf{z},x_j)d(\mathbf{z}',x_j)\ \forall \mathbf{z},\mathbf{z}' \in \mathcal{X} \tag{8}$$

An interesting property of the kernel matrix is that it is symmetric and positive semi-definite [18]. This characteristic will help to define a convex quadratic loss function for the clustering algorithm that can be optimized efficiently. Obviously, a clustering based on kernels can be extended easily to work with a given dissimilarity just considering the definition (8) for the kernel.

C. The Idealized Kernel of Dissimilarities

Let $\{x_i\}_{i=1}^n \in \Re^d$ be a set of objects. We are given weak supervised information to learn the distance metric in the form of similarity/ dissimilarity constraints. Must link constraints provide pairs of objects that are considered similar and cannot link constraints identify dissimilar ones. Let S and D be the subset of object pairs known to be similar/dissimilar. Mathematically they are defined as:

$$\mathcal{S} = \{(x_i, x_j) : x_i \text{ is similar to } x_j\} \tag{9}$$

$$\mathcal{D} = \{(x_i, x_j) : x_i \text{ is dissimilar to } x_j\} \tag{10}$$

Next, the idealized kernel is defined with the aim of maximizing the separability among different clusters. First, notice that the kernel function is a dot product in feature space $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ [18]. Therefore, it can be considered a similarity measure defined in the reproducing kernel Hilbert space [20]. For clustering applications, the ideal similarity (kernel) should be large for pairs of similar objects and small for dissimilar ones. Mathematically, the idealized kernel is defined for a family of kernels $\{k^l\}_{l=1}^M$ and a set of pairwise constraints (S, D) as follows:

$$K_{i,j}^* = k^*(x_i, x_j) = \begin{cases} max_l\{K_{ij}^l\} & \text{If } (x_i, x_j) \in \mathcal{S} \\ min_l\{K_{ij}^l\} & \text{If } (x_i, x_j) \in \mathcal{D} \end{cases} \tag{11}$$

Now, the idealized dissimilarity between two objects (x_i, x_j) is the euclidean distance in the kernel feature space induced by $\mathbf{k}^*$. Substituting the dot products by the idealized kernel, it can be written as:

$$\begin{aligned} d^2(x_i, x_j) &= \|\phi(x_i) - \phi(x_j)\|^2 \\ &= k^*(x_i, x_i) + k^*(x_j, x_j) - 2k^*(x_i, x_j) \end{aligned} \tag{12}$$

For pairs of similar objects the idealized dissimilarity takes the smallest value of the family of dissimilarities while for dissimilar ones takes the largest value. This measure will increase the cluster separability reducing the intra-cluster variance.

To illustrate the performance of the idealized dissimilarity let consider the breast cancer data set employed in the experimental section. We have applied a classical multidimensional scaling algorithm (MDS) [21] to project the data over a two dimensional subspace preserving approximately the original dissimilarities.

Fig. 1 shows the representation when the euclidean distance is considered and no supervisory information is available. The two classes (red-blue) overlap significantly and a clustering algorithm will fail to identify the two groups. Fig. 2 shows the projection for the MDS algorithm based on the idealized dissimilarity obtained from a family of 9 distances and a small set of randomly chosen pairwise constraints. We have considered 20% of all possible similarity/dissimilarity constraints. Similarity constraints are generated selecting pairs of patients that belong to the same class while dissimilarity constraints are retrieved from pairs of patients assigned to different classes. Fig. 2 shows that considering the idealized similarity both clusters become separable. Obviously, this measure may increase the overfitting. Therefore, the algorithm proposed to learn this dissimilarity should take care of this problem.

The idealized kernel (11) defined here for weak supervised clustering problems is related to the one proposed by [22] for classification:

$$k(x_i, x_j) = \begin{cases} 1 & if\ y_i = y_j \\ 0 & otherwise, \end{cases} \tag{13}$$

where y_i denotes the class label for x_i. However, the definition (13) takes into account only the class labels missing relevant information about the probability distribution for the objects. By contrast, the idealized kernel presented here takes into account a set of dissimilarity measures and hence, considers the probability distribution for the data. Besides, the kernel definition (13) is only valid for supervised problems in which class labels are available for the training set. It cannot be considered to incorporate incomplete knowledge in the form of equivalence constraints.

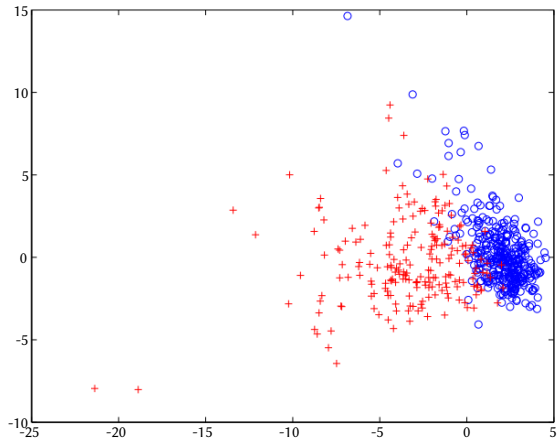


Fig. 1. Multidimensional scaling algorithm for a breast cancer dataset based on the euclidean distance. Both clusters (control and cancer) are quite overlapped in the projection.

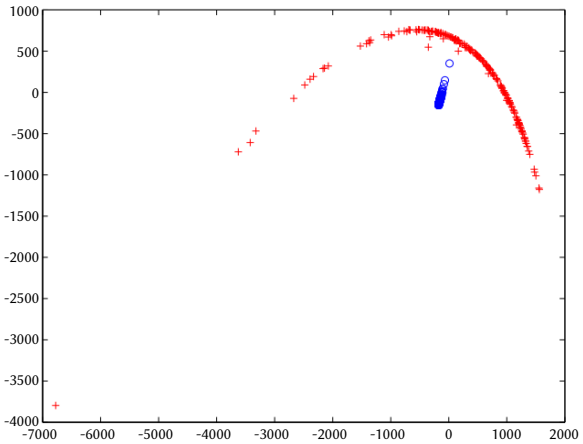


Fig. 2. Multidimensional scaling algorithm for a breast cancer dataset based on the idealized dissimilarity. Now, the two groups (control and cancer) can be easily identified by a clustering algorithm.

D. Multiple Kernel Learning for Clustering Algorithms Using Pairwise Constraints

In this section, we present the algorithm to learn the linear combination of similarities (kernels) that maximizes the cluster separability considering a set of pairwise constraints.

Let $\{d_{ij}^l\}_{l=1}^M$ be a set of M dissimilarity matrices that may come from different definitions or considering different features of the data. The dissimilarities are introduced into the clustering algorithm using the empirical kernel map (8). Let $\{k^l\}_{l=1}^M$ be the family of kernels obtained. Considering non-linear kernels will extend the original family of dissimilarities by non-linear mapping to a feature space. The problem can now be formulated as learning an optimal combination of kernels that maximizes the separability among the clusters using the pairwise constraints.

The linear combination of kernels is defined as:

$$k(x_i, x_j) = \sum_{l=1}^{M} \beta_l \, k^l(x_i, x_j) \tag{14}$$

where the β_l coefficients are constrained to be ≥ 0. Therefore, provided that each kernel is symmetric and positive semi-definite, the linear combination of kernels with $\beta_l \geq 0$ will be convex and positive semi-definite [23]. This property will help to define a convex quadratic loss function for the distance learning algorithm that may be optimized efficiently. Linear combination of kernels are preferred in this research over non-linear ones [4] because they are more robust to overfitting and the estimation of the parameters is more efficient computationally. The β_l coefficients are learned considering that he linear combination of kernels (14) should approximate the idealized kernel (11) with minimum error subject to the similarity/dissimilarity constraints. This optimization problem can be formulated in the primal as follows:

$$\min_{\beta,\xi} \Omega(\beta) + \frac{C_S}{N_S} \sum_{(x_i,x_j) \in S} \xi_{ij} + \frac{C_D}{N_D} \sum_{(x_i,x_j) \in D} \xi_{ij} \tag{15}$$

$$\text{s.t.} \quad \beta^T \mathbf{K}_{ij} \geq K^*_{ij} - \xi_{ij} \quad \forall (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \tag{16}$$

$$\beta^T \mathbf{K}_{ij} \leq K^*_{ij} - \xi_{ij} \quad \forall (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \tag{17}$$

$$\beta_l \geq 0 \quad \xi_{ij} \geq 0 \quad \forall l = 1, \ldots, M \tag{18}$$

C_S and C_D are regularization parameters that penalize training errors in the estimation of the idealized kernel. Particularly, they penalize similarity/dissimilarity constraint violations respectively. Both parameters may be determined by a grid search strategy using ten fold-crossvalidation. N_S, N_D are the number of pairwise constraints in S and D. $\Omega(\beta)$ is a regularization function that penalizes the complexity of the linear combination of kernels learned. Increasing the values of the regularization parameters C_S and C_D will minimize training errors in the constraints satisfaction but will increase the complexity of the similarity/kernel learned and the overfitting of the data. K^*_{ij} is the idealized kernel matrix introduced in section C and ξ_{ij} are the slack variables which are greater than zero for errors in the constraints satisfaction. Finally, K_{ij} is a matrix defined as $K_{ij} = \left[K^1_{ij}, K^2_{ij}, \ldots, K^M_{ij}\right]^T$, where K^l_{ij} is the idealized kernel matrix for similarity l.

The equations (16)-(17) model the constraints and ensure that the combination of similarities/ kernels learned are $\geq K^*_{ij}$ for similar objects and $\leq K^*_{ij}$ for dissimilar ones. The choice of the functional regularization term $\Omega(\beta)$ will determine the properties of the solution obtained. The L_1 norm is frequently considered in the Multiple Kernel Learning (MKL) literature [16], [24]. In this case, the solution will become sparse [25] and only a small set of similarities/kernels correlated with the idealized similarity will have non-null coefficient. However, in the bioinformatics applications considered in this paper, we are given frequently a small set of curated dissimilarities coming from different sources or distance metric definitions. Sparse solutions may lose relevant information and worsen the clustering results obtained [25].

Another choice for the regularization function $\Omega(\beta)$ is the L_2 norm. This penalization term distributes the weights more evenly reducing the value of the coefficients for less relevant kernels without removing them. Some authors have suggested in the literature that the L_2 norm gives better results in biomedical applications [25], [26]. Therefore, in this paper we will consider the L_2 norm regularization function.

Substituting $\Omega(\beta)$ by the L_2 norm the optimization problem in the primal is now formulated as follows:

$$\begin{aligned} \min_{\beta,\xi} \quad & \frac{1}{2} \| \beta \|^2 + \frac{C_S}{N_S} \sum_{(x_i,x_j) \in S} \xi_{ij} + \frac{C_D}{N_D} \sum_{(x_i,x_j) \in D} \xi_{ij} \\ \text{s.t.} \quad & \beta^T \mathbf{K}_{ij} \geq K^*_{ij} - \xi_{ij} \; \forall (\mathbf{x}_i \; \mathbf{x}_j) \in \mathcal{S} \\ & \beta^T \mathbf{K}_{ij} \leq K^*_{ij} + \xi_{ij} \; \forall (\mathbf{x}_i \; \mathbf{x}_j) \in \mathcal{D} \\ & \beta_l \geq 0 \; \xi_{ij} \geq 0 \; \forall l = 1, \ldots, M \end{aligned} \tag{19}$$

The previous constrained optimization problem can be solved using the method of Lagrange multipliers. Next, the problem can be written in the dual as follows

$$\begin{aligned} \max_{\alpha_{ij},\gamma} \quad & -\frac{1}{2} \sum_{\substack{(x_i,x_j) \in \mathcal{S} \\ (x_k,x_l) \in \mathcal{S}}} \alpha_{ij}\alpha_{kl}\mathbf{K}^T_{ij}\mathbf{K}_{kl} - \frac{1}{2} \sum_{\substack{(x_i,x_j) \in \mathcal{D} \\ (x_k,x_l) \in \mathcal{D}}} \alpha_{ij}\alpha_{kl}\mathbf{K}^T_{ij}\mathbf{K}_{kl} \\ & + \sum_{\substack{(x_i,x_j) \in \mathcal{S}, \\ (x_k,x_l) \in \mathcal{D}}} \alpha_{ij}\alpha_{kl}\mathbf{K}^T_{ij}\mathbf{K}_{kl} - \sum_{(x_i,x_j) \in \mathcal{S}} \alpha_{ij}\gamma^T\mathbf{K}_{ij} \\ & -\frac{1}{2}\gamma^T\gamma + \sum_{(x_i,x_j) \in \mathcal{D}} \alpha_{ij}\gamma^T\mathbf{K}_{ij} + \sum_{(x_i,x_j) \in \mathcal{S}} \alpha_{ij}K^*_{ij} \\ & - \sum_{(x_i,x_j) \in \mathcal{D}} \alpha_{ij}K^*_{ij} \end{aligned}$$

subject to:

$$0 \leq \alpha_{ij} \leq \begin{cases} \frac{C_S}{N_S} & \text{for } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \\ \frac{C_D}{N_D} & \text{for } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \end{cases} \tag{20}$$

$$\gamma_l \geq 0 \quad \forall l = 1, \ldots, M \tag{21}$$

where α_{ij} and γ_l are the Lagrange multipliers. The optimization problem in the dual is convex and quadratic without local minima [27]. Besides, the computational burden depends on the number of active constraints, that is those with $\xi_{ij} \geq 0$. This is more efficient than solving the problem in the primal where the computational complexity is proportional to the number of variables.

Once the α_{ij} and γ_l are estimated in the dual, the coefficients β_l for the linear combination of kernels can be obtained from $\frac{\partial L}{\partial \beta} = 0$. The vector of coefficients can be written as:

$$\beta = \sum_{(x_i,x_j) \in S} \alpha_{ij} \mathbf{K}_{ij} - \sum_{(x_i,x_j) \in D} \alpha_{ij} \mathbf{K}_{ij} + \gamma \tag{22}$$

Substituting in equation (14) we obtain the optimal combination of kernels learned from a set of equivalence constraints. Then, any clustering algorithm that works directly from a kernel matrix may be extended to incorporate a linear combination of dissimilarities. This will help to identify clusters that are non-separable using a single metric.

E. Support Vectors and KKT Complementary Conditions

In this section we study the relation between the value of the Lagrange multipliers and the constraints satisfaction. We also comment the meaning of the support vectors in the context of Multiple Kernel Learning.

The value of the Lagrange Multipliers α_{ij} determines if the linear combination of kernels complies with the constraints (16)-(17). To study this relation more in depth, let consider the Karush-Kuhn-Tucker (KKT) complementary conditions [27] for the optimization problem (19). They can be written in the primal as follows:

$$\begin{aligned} & \alpha_{ij} \left(\beta^T \mathbf{K}_{ij} - K^*_{ij} + \xi_{ij}\right) = 0 && (x_i, x_j) \in \mathcal{S} \\ & \alpha_{ij} \left(\beta^T \mathbf{K}_{ij} - K^*_{ij} - \xi_{ij}\right) = 0 && (x_i, x_j) \in \mathcal{D} \\ & \eta_{ij}\xi_{ij} = 0 && (x_i, x_j) \in \mathcal{S}, \mathcal{D} \\ & \gamma_l\beta_l = 0 && \forall_l = 1, \ldots, M \end{aligned} \tag{23}$$

From the previous KKT complementary conditions the following properties can be derived:

For similarity constraints, that is pairs of $(x_i, x_j) \in S$

$$\beta^T \mathbf{K}_{ij} = \begin{cases} = K^*_{ij} & 0 < \alpha_{ij} < \frac{C_S}{N_S} \\ > K^*_{ij} & \alpha_{ij} = 0 \\ < K^*_{ij} & \alpha_{ij} = \frac{C_S}{N_S} \end{cases}$$

For dissimilarity constraints, that is pairs of $(x_i, x_j) \in D$

$$\beta^T \mathbf{K}_{ij} = \begin{cases} = K^*_{ij} & 0 < \alpha_{ij} < \frac{C_D}{N_D} \\ < K^*_{ij} & \alpha_{ij} = 0 \\ > K^*_{ij} & \alpha_{ij} = \frac{C_D}{N_D} \end{cases}$$

The above properties show that when the similarity/dissimilarity constraints are satisfied with a margin larger than zero, the Lagrange multipliers α_{ij} are null and the corresponding similarity for the pair of objects will not appear in the solution. On the other hand, when the linear combination of kernels fails to satisfy the constraints or they are satisfied with margin exactly equal to zero the Lagrange multipliers are non-null and the similarity for the corresponding pair of objects will be considered in the solution. They are the support vectors and the optimization problem can be formulated exclusively in terms of them. Therefore, the complexity of the optimization algorithm will not depend on the size of the training set but on the number of the support vectors.

F. Datasets Description

We have considered a wide range of data sets to check the performance of the clustering algorithm proposed. Table I shows the different datasets considered and their features. The first three rows correspond to benchmark datasets retrieved from the UCI machine learning database (http://archive.ics.uci.edu/ml/datasets/). The last two rows are complex bioinformatics problems aimed to identify human cancer samples using the gene expression profiles. Both datasets can be recovered from a public webpage (http://bioinformatics2.pitt.edu). We have selected applications with wide range of signal to noise ratio (Var./Samp). In particular the cancer datasets (last two rows) have a high signal to noise ratio with large number of variables and small number of samples. Therefore, they are problems that favor the overfitting of the data and will serve to check the generalization ability of the algorithm proposed. Moreover, as the number of samples is small, the supervisory information available is also quite limited and learning the metric becomes a challenging task. For all the datasets the class label is available. This will help to evaluate rigorously the clustering results considering objective measures. Finally, the variables have been normalized subtracting the median and dividing by the inter-quantile range.

TABLE I. Properties of the Different Data Sets Considered

Data sets	Samples	Variables	Var./Samp	Classes
Wine (UCI)	177	13	0.17	3
Ionosphere (UCI)	351	35	0.01	2
Breast Cancer (UCI)	569	32	0.05	2
Lymphoma	96	4026	41.9	2
Colon Cancer	62	2000	32	2

III. Results

In this section we first comment the preprocessing of the datasets and how the supervisory information is generated. Next, the set of dissimilarities considered by the learning algorithm are introduced as well as the method to estimate the parameters. Finally, we describe the objective measures to evaluate the clustering algorithms and the experimental results are discussed.

Cancer samples using the gene expression profiles are represented in high dimensional spaces with high level of noise to signal ratio. Noisy features may deteriorate the clustering performance. Therefore, feature selection to remove redundant variables is recommended to improve the clustering results [10]. To this aim, genes (features) are ranked by the interquantile range (IQR). Those genes with small variability are considered irrelevant to discriminate between different disease states. We have considered five subsets with the 280; 146; 101; 56 and 34 genes ranked higher considering the IQR. Supervised feature selection algorithms are not considered because in clustering problems class labels are not available. For clustering algorithms based on a single dissimilarity we have chosen the subset of genes that gives rise to the smallest error. Clustering methods based on multiple kernels consider all the dissimilarity matrices obtained from different subset of features. It is expected that the learning algorithm will help to remove dissimilarities based on noisy features.

Regarding the set of dissimilarities integrated into the clustering algorithm we have considered nine measures widely used in bioinformatics applications. Euclidean, Manhattan, Chevichev, Mahalanobis, Cosine, Correlation, Spearman, Kendall-τ and χ^2. In order to build an ensemble of dissimilarities we have considered for each distance different subsets of features and non-linear transformations using kernel methods. After that, we obtained an ensemble of 45 dissimilarity matrices for each type of kernel.

To generate the set of pairwise constraints we have followed the approach of [13]. The similarity constraints S are obtained by sampling randomly all the object pairs that belong to the same class. The size of S is chosen such that the number of connected components is approximately the 20% of the number of objects. The dissimilarity constraints D are chosen sampling randomly the object pairs that belong to different classes. Twenty independent random sets for S and D are generated and the average error is reported.

The optimal values for the regularization parameters C_S and C_D are estimated using a grid search strategy and the errors are computed by ten-fold cross-validation over the set of constraints. The number of clusters for each problem has been set up to the number of classes. As kernel k-means algorithm is sensitive to the initialization we have reported the average error over 20 independent trials with random initialization.

The clustering algorithms have been evaluated by two error functions widely used in the literature [13]. The first one is the accuracy. It determines the probability that two objects that belong to the same or different classes are grouped in the same way by the clustering algorithm. Mathematically it can be defined as:

$$\text{accuracy} = \sum_{i>j} \frac{1\{1\{y_i = y_j\} = 1\{\hat{c}_i = \hat{c}_j\}\}}{0.5\, N\,(N-1)} \tag{24}$$

where y_i is the reference class label for object i and c_i is the group assigned to object i by the clustering algorithm. N is the number of objects in the dataset. The accuracy may lead often to wrong conclusions because the average value for two random partitions is greater than zero. To overcome this problem, it has been proposed in the literature the adjusted randindex [28].

Table II and Table III compare the different clustering algorithms according to accuracy and the adjusted randindex. First row provides the results for the semi-supervised learning algorithm proposed in this paper and based on an ensemble of dissimilarities. Polynomial kernels allow to increase the number of dissimilarities incorporating non-linear transformations of the original ones. We have compared in the second row with a standard clustering method that learns the metric from pairwise constraints [13]. Third row provides the performance of kernel k-means based on the best measure for the whole family of dissimilarities considered. Each column reports the best distance regarding the data set analyzed. Again, the original dissimilarities may be non-linearly transformed to obtain more general measures using polynomial kernels. Finally, last row shows the results for k-means standard clustering algorithm based on the euclidean distance. Polynomial kernels allow us to transform non-linearly this metric to consider more general dissimilarities and non-spherical groups.

TABLE II. Accuracy for the Semi-supervised Clustering Algorithm Proposed Versus Other Approaches. The Results Are Averaged Over Twenty Independent Random Subsets for S and D

Technique	Kernel	Wine	Ionosphere	Breast	Colon	Lymphoma
Clustering proposed	Linear	0.94	0.90	0.92	0.89	0.95
	Pol. 3	0.96	0.89	0.92	0.90	0.92
Metric learning (Xing)	Linear	0.87	0.74	0.85	0.87	0.90
	Pol. 3	0.51	0.74	0.86	0.88	0.90
Kernel K-means (Best dissimilarity)	Linear	0.94	0.88	0.90	0.88	0.94
	Pol. 3	0.94	0.88	0.90	0.88	0.93
		$\chi 2$	*Mahalanobis*	*Manhattan*	*Correlation*	$\chi 2$
K-means (Euclidean)	Linear	0.92	0.72	0.88	0.87	0.90
	Pol. 3	0.87	0.73	0.88	0.88	0.90

TABLE III. Adjusted RandIndex for the Semi-Supervised Clustering Algorithm Proposed Versus Other Approaches. The Results Are Averaged Over Twenty Independent Random Subsets for S and D

Technique	Kernel	Wine	Ionosphere	Breast	Colon	Lymphoma
Clustering proposed	Linear	0.82	0.63	0.69	0.60	0.79
	Pol. 3	0.85	0.60	0.69	0.63	0.73
Metric learning (Xing)	Linear	0.68	0.23	0.50	0.54	0.66
	Pol. 3	0.50	0.23	0.52	0.58	0.65
Kernel K-means (Best dissimilarity)	Linear	0.82	0.58	0.66	0.59	0.77
	Pol. 3	0.81	0.58	0.66	0.59	0.76
		$\chi 2$	*Mahalanobis*	*Manhattan*	*Correlation*	$\chi 2$
K-means (Euclidean)	Linear	0.79	0.20	0.59	0.59	0.65
	Pol. 3	0.67	0.21	0.60	0.59	0.65

From the analysis of Table II and Table III, we report three relevant conclusions:

First, the semi-supervised clustering algorithm proposed in this paper improves significantly the performance of a benchmark clustering algorithm developed by Xing [13] that learns the metric from pairwise constraints. The accuracy and the adjusted randindex are significantly improved even for the cancer datasets, with high level of noise and large number of variables. This result can be explained because the clustering proposed here has smaller number of parameters and the regularization term helps to reduce the overfitting. Notice also that our model integrates dissimilarities based on different sets of features removing the problem of choosing the optimal set of variables, which is a complex task in clustering problems.

To determine if the differences between our clustering algorithm and the one proposed by Xing are statistically significant we have computed the boxplots for both techniques. To this aim, we have generated 20 independent random sets of constraints for S and D and we have estimated the accuracy and the adjusted randindex. Fig. 3 shows the boxplots for the accuracy and Fig. 4 for the adjusted randindex. Odd numbers in the x-axis correspond to the boxplots for our clustering algorithm and the different datasets considered in the same order as in Table II. Similarly, even numbers correspond to the supervised clustering algorithm proposed by Xing. The boxplots show that the differences between both algorithms are statistically significant at 95% confidence level for all the datasets considered in this paper.

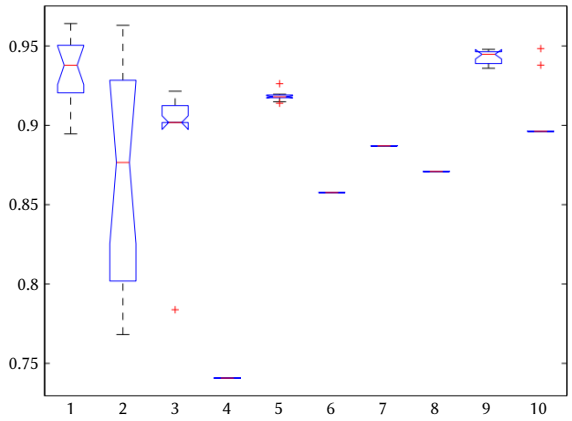


Fig. 3. Accuracy boxplots that compare the multiple kernel learning clustering proposed with the metric learning algorithm developed by Xing. 20 independent trials have been recorded considering 20 sets of constraints generated randomly.

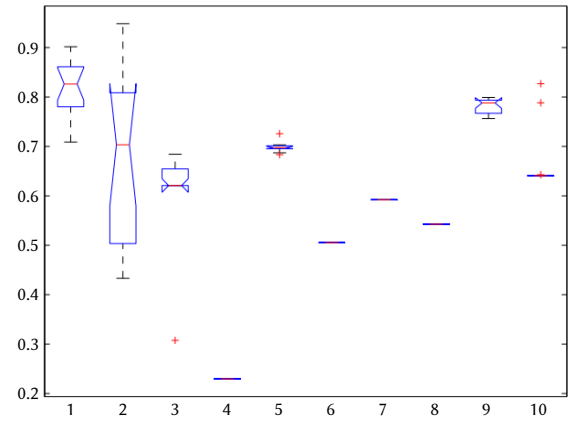


Fig. 4. Adjusted RandIndex boxplots. They compare the multiple kernel learning clustering proposed with the metric learning algorithm developed by Xing. 20 independent trials have been recorded considering 20 sets of constraints generated randomly.

The clustering algorithm proposed that integrates an ensemble of dissimilarities improves the accuracy and adjusted randindex of kernel k-means based on the best similarity. The combination of dissimilarities provides more information that a single measure. Besides, Table II and Table III show that the best dissimilarity depends on the particular problem considered. Moreover, for unsupervised applications choosing the best measure is an ill-posed problem, because no supervised index error can be defined to guide the selection of an appropriate metric. Our algorithm helps to overcome the problem of choosing an optimal dissimilarity, the best kernel and even the optimal subset of features. This is frequently a challenging task, for instance in complex bioinformatics applications.

Finally, we remark that the learning algorithm proposed improves significantly the standard k-means clustering algorithm based on the euclidean distance for all the datasets considered. The results are similar for a non-linear transformation of the euclidean distance considering polynomial kernels of degree 3.

IV. Discussion

Several algorithms developed in the literature to learn the distance metric are related to the one proposed here. The first approach tries to learn a full or diagonal Mahalanobis distance considering pairwise constraints [1], [12], [13]. Some authors have extended the previous techniques to more general dissimilarities using kernel methods [14], [15], [29]. However, they are based on a single distance metric that may fail to reflect accurately the objects proximities. Besides, as the number of parameters grows with the space dimensionality they are prone to overfitting and the computational complexity is high. Although new algorithms have been proposed to improve the computational efficiency and to reduce the overfitting [6] they suffer from similar drawbacks. Several differences are worth to mention with the approach proposed here. First our algorithm is able to integrate a set of dissimilarities that may exhibit different properties from a set of pairwise constraints. Second, the loss function incorporates a penalty term and has a small number of parameters which helps to reduce the overfitting. Finally, the optimization problem is quadratic, the complexity depends on the number of support vectors and it is efficient computationally.

Our approach is more related to multiple kernel clustering methods [7]–[9], [16] that are able to integrate different dissimilarities that come from different features or representations of the objects using kernel methods. However, these algorithms differ from our approach because they integrate the dissimilarities in an unsupervised way and the resulting metric may not help to improve the clustering results. In this way, some researchers have mentioned that learning the metric without any supervised information may be an ill-defined problem [15].

Finally, few authors have addressed the problem of multiple kernel learning from a set of pairwise constraints for clustering applications [17], [30]. However, they rely on complex optimization problems that are more difficult to solve than the one proposed in this research.

V. Conclusion

In this paper we have developed a semi-supervised learning algorithm to integrate an ensemble of kernels (similarities) into a clustering algorithm using weak supervision in the form of pairwise constraints. Our method offers three advantages over previous metric learning algorithms. First, it learns a combination of dissimilarities that may come from different features of the objects or different kernels. This strategy avoids the problem of choosing the right kernel (similarity), the best subset of features or the optimal value for the kernel parameters that may be a challenging task for certain type of applications. Second, the loss function is convex and quadratic and it may be efficiently optimized. Finally, the learning algorithm is robust to overfitting.

The clustering algorithm proposed has been applied to three benchmark datasets and to complex cancer identification problems based on the gene expression profiles. The experimental results suggest that learning a combination of similarities (kernels) improves the performance of a clustering algorithm based on the best similarity (kernel). Besides, the algorithm developed outperforms a standard semi-supervised clustering proposed in the literature that learns the metric from the data. In particular, our method performs significantly better for cancer problems with high level of noise to signal ratio which suggests that it is robust to overfitting.

Future research trends will focus on the application of this formalism to other bioinformatics problems such as gene function prediction.

References

[1] C. Shen, J. Kim, L. Wang, "Scalable large-margin mahalanobis distance metric learning," IEEE transactions on Neural Networks, vol. 21, no. 9, pp. 1524–1530, 2010.

[2] H. Fyad, F. Barigou, K. Bouamrane, "An Experimental Study on Microarray Expression Data from Plants under Salt Stress by using Clustering Methods," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 2, pp. 38-47, 2020.

[3] M. Martín-Merino, Á. Blanco, "A local semi-supervised sammon algorithm for textual data visualization," Journal of Intelligent Information Systems, vol. 33, no. 1, pp. 23–40, 2009.

[4] A. Woznica, A. Kalousis, M. Hilario, "Learning to combine distances for complex representations," in Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 1031–1038.

[5] A. Seal, A. Karlekar, O. Krejcar, E. Herrera-Viedma, "Performance and convergence analysis of modified C-means using jeffreys-divergence for clustering", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 2, pp. 141-149, 2021.

[6] B. Nguyen, B. De Baets, "Kernel-based distance metric learning for supervised k-means clustering," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 10, pp. 3084–3095, 2019.

[7] A. Huang, T. Zhao, C.-W. Lin, "Multi-view data fusion oriented clustering via nuclear norm minimization," IEEE Transactions on Image Processing, vol. 29, pp. 9600–9613, 2020.

[8] B. Zhao, J. T. Kwok, C. Zhang, "Multiple kernel clustering," in Proceedings of the 2009 SIAM International Conference on Data Mining, 2009, pp. 638–649, SIAM.

[9] J. Hu, M. Li, E. Zhu, S. Wang, X. Liu, Y. Zhai, "Consensus multiple kernel k-means clustering with late fusion alignment and matrix-induced regularization," IEEE Access, vol. 7, pp. 136322–136331, 2019.

[10] D. Huang, W. Pan, "Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data," Bioinformatics, vol. 22, no. 10, pp. 1259–1268, 2006.

[11] H. Zeng, Y.-m. Cheung, "Semi-supervised maximum margin clustering with pairwise constraints," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 5, pp. 926–939, 2011.

[12] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall, G. Ridgeway, "Learning a mahalanobis metric from equivalence constraints," Journal of Machine Learning Research, vol. 6, no. 6, 2005.

[13] E. Xing, A. Y. Ng, M. I. Jordan, S. J. Russell, "Distance metric learning with application to clustering with side-information," in Advances in Neural Information Processing Systems, vol. 15, 2002, MIT Press.

[14] B. Nguyen, B. De Baets, "Kernel distance metric learning using pairwise constraints for person reidentification," IEEE Transactions on Image Processing, vol. 28, no. 2, pp. 589–600, 2018.

[15] D.-Y. Yeung, H. Chang, "A kernel approach for semisupervised metric learning," IEEE Transactions on Neural Networks, vol. 18, no. 1, pp. 141–149, 2007.

[16] Y. Yao, Y. Li, B. Jiang, H. Chen, "Multiple kernel kmeans clustering by selecting representative kernels," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4983–4996, 2020.

[17] T. Yang, R. Jin, A. K. Jain, "Learning kernel combination from noisy pairwise constraints," in 2012 IEEE Statistical Signal Processing Workshop (SSP), 2012, pp. 752–755, IEEE.

[18] V. Vapnik, Statistical Learning Theory. New York: John Wiley & Sons, 1998.

[19] E. Pekalska, P. Paclick, R. Duin, "A generalized kernel approach to dissimilarity-based classification," Journal of Machine Learning Research, vol. 2, pp. 175–211, 2001.

[20] G. Wu, E. Y. Chang, N. Panda, "Formulating distance functions via the kernel trick," in Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 2005, pp. 703–709.

[21] M. Martín-Merino, A. Blanco, J. De Las Rivas, "Combining dissimilarities in a hyper reproducing kernel hilbert space for complex human cancer prediction," Journal of Biomedicine and Biotechnology, vol. 2009, 2009.
[22] N. Cristianini, J. Kandola, J. Elisseeff, A. Shawe-Taylor, "On the kernel target alignment," Journal of Machine Learning Research, vol. 1, pp. 1–31, 2002.
[23] C. Soon Ong, A. Smola, R. Williamson, "Learning the kernel with hyperkernels," Journal of Machine Learning Research, vol. 6, pp. 1043–1071, 2005.
[24] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, "Simple multiple kernel learning," Journal of Machine Learning Research, vol. 9, pp. 2491–2521, 2008.
[25] S. Yu, T. Falck, A. Daemen, L.-C. Tranchevent, J. A. Suykens, B. De Moor, Y. Moreau, "L2-norm multiple kernel learning and its application to biomedical data fusion," BMC bioinformatics, vol. 11, no. 1, pp. 1–24, 2010.
[26] M. Kloft, U. Brefeld, P. Laskov, K.-R. Müller, A. Zien, S. Sonnenburg, "Efficient and accurate lp-norm multiple kernel learning," in Advances in Neural Information Processing Systems, vol. 22, 2009, Curran Associates, Inc.
[27] W. Kaplan, Maxima and minima with applications: practical optimization and duality, vol. 51. John Wiley & Sons, 1998.
[28] L. Hubert, P. Arabie, "Comparing partitions," journal of classification, vol. 2, pp. 193–228, 1985.
[29] M. S. Baghshah, S. B. Shouraki, "Non-linear metric learning using pairwise similarity and dissimilarity constraints and the geometrical structure of data," Pattern Recognition, vol. 43, no. 8, pp. 2982–2992, 2010.
[30] B. Yan, C. Domeniconi, "An adaptive kernel method for semi-supervised clustering," in European Conference on Machine Learning, 2006, pp. 521–532, Springer.

Manuel Martín Merino

Manuel Martín Merino received the B. S. degree in physics from the University of Salamanca (Spain) in 1996 and the PhD. degree in applied physics from the same University in 2003. He is currently professor of Artificial Intelligence and supervisor of the Telefonica Chair in the Computer Science school at the University Pontificia of Salamanca. His research interests include Machine Learning, Artificial Neural Networks, Kernel methods, Support Vector Machines, clustering algorithms and their applications to Bioinformatics and text mining problems.

Alfonso José López Rivero

Alfonso José López Rivero. PhD in Computer Science. He is a professor, since 1996, and member of the research group GESTICON (Ethical and Technological Management of Knowledge) at the Pontifical University of Salamanca, UPSA, (Spain). He is a member of the organizing and scientific committee of several international symposia and co-author of articles published in several recognized journals, workshops and symposia. He has been Dean of the Faculty of Computer Science and Director of the Office for the Transfer of Research Results at UPSA.

Vidal Alonso Secades

Vidal Alonso was born in Luanco, Spain, in 1966. He received the Computer Science Degree in 1992 from the Polytechnic University of Madrid, and the Ph. D. degree, in 2004 from the Pontifical University of Salamanca, Spain. He was a Full Professor of Computer Science at the Pontifical University of Salamanca since 1994. He has occupied the position of Vice rector at his University for five years, until 2015. He also was the Director of the Computer Science School for nine years (2000-2009.) He works in data structures, knowledge discovery and data quality. Dr. Alonso is a member of ALI (Computer Science Spanish Association) and he won the Castilla y Leon Digital Award in 2007.

Marcelo Vallejo García

PhD in Computer Science. Bachelor in Business Administration. Currently Professor of Financial Economics and Accounting, in the Faculty of Computer Science of the Pontifical University of Salamanca. Among his former positions, we can highlight the following: Director of the Doctoral Program in Insurance, Legal and Business Sciences by Pontifical University of the Pontifical University of Salamanca, Vice Dean of the UPSA Faculty of Science Computer during the period 2015-2021 and Responsible of the module "Union Economic and Monetary", financed by the European Commission, as a part of Jean Monnet actions. With more than 30 years of teaching experience, he is the current coordinator of Bachelor's Degree in Technology Business of the Pontifical University of Salamanca. He has participated as a collaborating researcher and principal investigator in several competitive projects related to their areas of research and teaching. He is the author and co-author of scientific publications and has participated as speaker at several national and international conferences, mainly in topics relatives to corporate reputation and management and reporting of financial and no-financial business information.

Antonio Ferreras García

Antonio Ferreras, received a PhD in Telecommunications Engineering, in 1995 from the Polytechnic University of Madrid for his studies in Optoelectronics. He also has degrees in Economics, Psychology and Law: from 2018 he is professor at the Pontificial University of Salamanca. Currently, his research area is focused on cybersecurity and data science.

Normative Affordances Through and By Technology: Technological Mediation and Human Enhancement

Niklas Alexander Döbler[1,2,3*], Clemens Bartnik[4]

[1] Department of General Psychology and Methodology, University of Bamberg, Bamberg (Germany)
[2] Research Group EPÆG (Ergonomics, Psychological Aesthetics, Gestalt), Bamberg (Germany)
[3] Bamberg Graduate School of Affective and Cognitive Sciences (BaGrACS), University of Bamberg, Bamberg, (Germany)
[4] Video & Image Sense Lab, Institute of Informatics, University of Amsterdam, Amsterdam (The Netherlands)

Received 5 April 2022 | Accepted 29 June 2022 | Early Access 19 September 2022

unir
LA UNIVERSIDAD EN INTERNET

Abstract

Human activity is fundamentally embedded in and constituted by technology. In this regard, technology influences not only how people experience the world, but also which possibilities for action offered by the environment (affordances) can be perceived and ultimately acted upon. As having socio-cultural and normative aspects, affordances are deeply relational to the technological human form of life. Postphenomenology describes several human-technology relations and their perception and action mediating effects. Therefore, it provides a suitable framework to examine how technology mediates the perception of affordances and leads to different behavioral outcomes. Technology can reveal hitherto hidden affordances but can also result in the manipulation and concealment of action possibilities. Both aspects can be deliberately controlled by using a particular technology and/or interfering with the technological hermeneutic process. Technological mal-functions, limitations, purposeful corruption, or human error can disrupt the hermeneutic qualities of technology and may lead to false conclusions about affordances and respective maladaptive behavioral outcomes. Technology can also be applied to humans to form "better" versions of them. One consequence of these so-called Human Enhancement technologies is the emergence of different affordances for the enhanced individual and the possible establishment of new affordances inside a form of life. Manipulating the perception and emergence of affordances through technological mediation or Human Enhancement can have severe political and ethical consequences. It is necessary to engage in an open debate about the perception and action mediating power of technology and the human reliance on them in our current and future form of life.

Keywords

Affordances, Ethics, Human Enhancement, Postphenomenology, Technology.

DOI: 10.9781/ijimai.2022.09.006

I. Introduction

Throughout their existence, members of the genus Homo - exploiting the rich landscape of material opportunities - have constantly altered the face of the Earth. From the first combination of different materials for the purpose of constructing composite tools to today's particle accelerators: Human activity cannot be separated from its embodiment through and with technical means [1]. The presence of technology not only shapes activity in the human lifeworld, but also influences human perception at various levels (individual/cultural), in a way that "is more than a formal change; the way world is experienced is changed ontologically" [1, p. 47].

Concerning an animal's perception of its environment, Gibson [2] states that its perceptual system is optimized to process visual features that enable ecologically important behaviors. Subsequently, he defines the concept of *affordances*. Initially, affordances describe what an environment offers to an animal. In the following scientific discussion, the concept was extended to take human activity in the material world into account [3]. Nowadays, affordances are conceived as possibilities for action and, in the case of humans, these possibilities are deeply embedded in socio-cultural and socio-material relations [4]–[6] and thereby connected to the notion of Wittgenstein's *form of life*:

> Affordances are possibilities for action the environment offers to a form of life, and an ecological niche is a network of interrelated affordances available in a particular form of life on the basis of the abilities manifested in its practices—its stable ways of doing things. [5, p. 330]

Imagine a form of life as a set of lived practices and available common behavioral patterns. It structures and enables our activities but simultaneously consists of these aspects. It "lives in use" [7]. Using technologies is a stable and regular way of human's meaningful engagement with the world, reciprocally and skillfully shaping it and the possibilities of action afforded by it [1],[7]–[11]. Hence, technologies and their normative use are part of the human form of life, that shapes and structures our interactions with existing and

* Corresponding author.
E-mail address: niklas.doebler@uni-bamberg.de

future technologies [7]. Most importantly, the ability to make a correct epistemic judgment is also part of a form of life [12].

Perception and realization of affordances depend on the practices and abilities in a form of life [5]. Technology in its use actively shapes human perception and thus influences which affordances can emerge and are realized in the human form of life. Hence, technologies are part of the human ecological niche; the way we live [2],[5]. In other words: By influencing the perception of affordances, technology influences which patterns of behavior emerge from these possibilities, become widely available, and eventually manifest themselves in concrete practice. What is important here is that adopting these behavioral patterns is not just a matter of social persuasion and convention but begins much earlier: In the mediated perception and notion of what is possible in the first place.

Given the ubiquitous influence of human technology and our general dependence on it for survival, the technological form of life comes with an array of ethical issues: "Technologies help to determine how people act, so that it is not only people but also things who give answers to the classical moral question, 'How to live?'" [13, p. 236]. Thus, the way human beings live simultaneously shapes moral behavior and the normative notion of certain standards [7],[14]. Furthermore, affordances relate to the value of technologies [15],[16]. The value is determined by the affordances, their possible realization, the user's intentions, and other contextual parameters [15]. The context here could be a form of life that partially determines which affordances can be realized in the first place. In addition, the significance of affordances arises from past experience and forward-looking expectations [17]. To be realized an affordance must relate to this specific form of life. The same form of life that produced the technology in question. This recursive process is why a form of life and all the possible technological engagements it incorporates can be understood as a "river-bed," as something that is in steady flux but at the same time provides structure and stability [7].

Given the close connection between affordances and everyday phenomenology [6], we must turn towards these factors that shape the regular perception of the human lifeworld. Most notably, this lifeworld is a technological one [1]. In it, the influence of technology on human perception and action is constitutive, enhancing, and mediating [1],[11],[13]. This means that human perception of affordances is deeply connected to the mediational processes of technology. Hence, the form of life and the available affordances depend on how technology alters human perception. To fully understand the consequences of different human-technology relations, we will use the approach of *postphenomenology* and its perspective on technological mediation of human experience to demonstrate how technology can change the perception of affordances and eventually a form of life.

II. Marrying Postphenomenology and Affordances

Postphenomenology is concerned with how technology influences the human experience of the lifeworld. Postphenomenological research is interested in the technological transformation and mediation of human experience, perception, and action and the related embodied perspective [1],[18]–[20].

According to Don Ihde [1], human beings and technology engage in different formalizable relationships [1],[20],[21]. Among others, one example are so-called hermeneutic relations, in which technology translates information about the world into a "text" that is understandable by human beings.

Hermeneutic relation: Human → (Technology – World)

Here the perceptual focus is *upon* the technology. Ihde employs a thermostat as an example here. By "reading" its "text" (units of temperature), we are able to perceive an aspect of the world, even without directly experiencing it. Note the first connection to a given form of life and its skills and abilities. The affordance of the thermostat in terms of reading and insight into the temperature requires knowledge of the relation of the present symbols (usually numbers), the measurement unit (usually degrees of Celsius or Fahrenheit), and the respective embodied experience of temperature. If one part of this tripartite relationship is unknown, we cannot make sense of the hermeneutic text. As most of the world measures temperature in Celsius, we also encounter the cultural and spatial aspects of the form of life. For someone unfamiliar with the mathematical relationship between Fahrenheit and Celsius, 50 °F is incomprehensible.

Additionally, Ihde describes the so-called alterity relation. When engaging with technology, artifacts, depending on their perceived automatism and interactional potential, are sometimes experienced as an "(Quasi)-Other," which steps to the foreground and becomes an interactional partner and the focus of the experience. But unlike the hermeneutic relationship, there may be no connection to the outside world at all, or the same world may withdraw to the background. The minus inside the parentheses formalizes this.

Alterity relation: Human → Technology – (–World)

Alterity relations are constituted by relating *to* technology in a specific way. An example is the deliberately anthropomorphized assistant, such as Apple's "Siri." When somebody asks this software about the weather, questioners relate to this technology as to somebody other who knows about the weather. This interactional component between a so perceived quasi-other and the human user distinguishes the alterity relation from the mere hermeneutic relation. However, demarcations between the different relations are complicated, and Ihde acknowledges the emergence of descriptive grey areas.

Phenomenology sees itself first and foremost as a movement that investigates the relationship between human beings and their lifeworld, rather than being a mere description of reality [13]. Postphenomenology seeks to investigate technological mediation "from within," using insights to shape intentions and actions [22]. As possibilities of action, affordances are fundamentally relational [5],[6],[15],[23],[24], especially towards a particular form of life [5] and individual skills and abilities [12]. Therefore, a postphenomenological perspective could shed light on human beings' relations with their environment and the actions afforded. This becomes evident, considering how affordances can contribute to the postphenomenological notion of multistability, meaning flexible but finite possibilities for using an artifact [25].

Postphenomenology is not only concerned with the mediation of perception but also with the mediation of action. Here technology is simultaneously inviting and inhibiting [13]. As affordances can be conceived as invitations for behavior [26], the important connection of this concept to technology and its perception and action mediating character becomes even clearer.

Using technology, humans relate to the world in different ways. These relations depend on the affordances perceived and the emerging actions. Furthermore, these relations and the resulting affordances shape our future engagement with the environment. This is the diachronic dimension of affordances: They reflect our past and future [17]. Therefore, the theory of affordances, combined with a postphenomenological perspective, can contribute to our understanding of how different technological mediations translate into different stable patterns of behavior and become a part of the human form of life.

III. An Example of the Perception of Affordances Through Technological Mediation

Imagine a vast and empty horizontal almost entirely bounded by impassable cliffs. The ground is covered with grass, and approximately 300 meters away from where your stand is a podium topped with a gold-filled pot. The only way to get there is a direct passage over the plane. Luckily, there are no physical obstacles. This situation now has various affordances: For someone seeking imminent monetary fortune, the most relevant one may be the possibility of crossing the plane and taking the gold. Unfortunately, the whole plane is heavily contaminated by gamma radiation. Attempting to cross it would be inevitably fatal for any human. With a wavelength of < 10 pm, gamma radiation is invisible to the human visual system. This is an example of hidden affordances [27], as the described landscape may prima facie afford safe passage to the pot of gold but, in fact, bears severe health danger. Note that the original conception of affordances was not only concerned with beneficial offers but also with maladaptive ones [2].

There are several possibilities for gaining knowledge about the hidden affordance of the plane. Once one person started crossing it and suffered from terminal radiation sickness, other observers could make assumptions concerning the hitherto imperceptible hazard and thus about the hidden affordances. However, they would not be able to tell what exactly caused the person's death, though they might a suspect a deadly invisible and mysterious hazard on the plane. At some point, they may (correctly) conclude that traversing the terrain is impossible and therefore refrain from further attempts. These conclusions and the resulting normative call to inhibit crossing behavior would be drawn through a tragic instance of observational learning.

Given the human communicational abilities, eye-witnesses of the deadly crossing attempt could tell other humans about the danger and thus influence their notion of affordances even without them perceiving the plane on their own. Cultural transmission preserves knowledge about the mysterious yet dangerous plane. Some foolhardy adventurers may try to cross it from time to time, but their attempts would always have the same fatal result. Here the normative dimension of affordances and epistemic judgments becomes evident again: The normative standards of how to engage with specific affordances are tied to individual skills and abilities in a given situation [12] but also to socio-cultural and -material practices and customs [5] (for a slightly different view, see [28]) and therefore a form of life [7]. The existence of such a form of life makes the practice possible in the first place [7],[29]. Hence, the ability to recognize the "correct" affordance of the deceptively safe plane is tied to socio-cultural transmission, shaping the expectations of the plane, which are then re-enacted in concrete practice other community members can draw from [6]. Accordingly, affordances cannot be integrated into a given affordance landscape without shared attention and a mutual understanding of the meaning and embodied experience.

The notion of expectations is crucial. In the realm of so-called "cultural affordances," expectations are tied to "conventional affordances," which require a correct inference by the perceiver [6]. The set of shared expectations creates a "local ontology" interwoven with concrete practices and socio-material reality [6]. Engagement with the material environment is an essential condition for the emergence of customs, practices, and meaning [10],[11]. If someone placed a warning sign at the edge of the plane, the normative aspect of the conventional affordance would also be communicated socio-materially. The meaning of this material sign is enactively embedded in past engagement with its content and modulation of attention [6],[11]. The social practice relating to the material sign can therefore help to define a set of local and relevant affordances from the whole available landscape of affordances [6]. Once sufficiently adopted and recognized, culturally transmitted information becomes a part of a specific form of life and would thus enable its sufficiently skilled members to detect relating affordance even without experiencing the radiation first hand.

Another means of assessing the hidden affordances of the plane is science. Here one must note that science is embodied in technology [1],[20], meaning that scientific insights heavily depend on the available technology that produces them and vice versa. Scientific observation is socio-materially augmented perception [30].

In our example, one way to learn about the hidden hazard on the plane would be to discover and formulate the physical principle of radioactivity and its maladaptive effect on biological tissue. But this is only the first step. Even if known that there is such thing as radioactivity, it is not clear if it is the cause of the plane's danger. One way to find out is technological aid, specifically a Geiger counter. Using this technological device constitutes a hermeneutic relationship with the plane. It allows for the perception of an environmental feature through technology and reveals a hitherto hidden affordance.

$$Human \rightarrow (Geiger\ counter - World)$$

The Geiger counter transforms the information in a way perceivable by humans. Feeling the adverse effect of the radiation is sufficient to refrain from trying to cross the plane but insufficient for understanding what causes this experience. Furthermore, the Geiger counter can deliver the relevant information before feeling the radiation directly. We can simply "read" it from a screen without exposing ourselves to lethal danger.

Affordances are always specific to the particular animal and depend on their bodily condition [2],[31]. This fact must be accounted for in the technologically mediated perception of the environment. As scientific instruments possess the ability to perceive what may be hidden from humans, the detections resulting from this "instrumental realism" [32] must be compatible with the human condition: "for embodied humans whose observations are those of bodily-perceptual creatures, […] the information, data, or image must be transformed, translated, into what is open to our *anthropological constant, an embodied human*" [32, p. 113, italics from the original]. In other words, the technologically retrieved information about the world must be presented in a sensory and cognitively comprehensible way to humans. Once again, converging with the concept of conventional affordances and their requirement of correct inference regarding certain expectations [6], the notion of instrumental realism requires scientists, engineers, and designers to think about the technological mediation processes within the scientific endeavor when integrating their findings in the expectations and predictions of their models and theories. Furthermore, they must acknowledge the emergence of a combined human and technological "composite intentionality" in the hermeneutic relationship between humans and technology [21].

In a recent study on the fMRI-supported neuropsychiatric diagnostic process, de Boer and colleagues showed how this imaging technology mediates researchers' notion of brain complexity and materializes deduced diagnostic labels in concrete experimental diagnostic practice [33]. The establishment of an ontological link between a scientific measurement and a diagnosis shapes expectations and materializes an affordance for future use. Regardless if this concerns a deadly dose of radiation or psychiatric diagnosis: What matters is the socio-material commitment of scientifically retrieved information with its alleged meaning by applying skillful inference and deduction. This is no arbitrary process. People will still suffer from radiation sickness, even without knowing what the Geiger counter display means, and observers may expect the same outcome. Importantly, any materialization of conventional affordances inherent to a local ontology must ensure the appropriateness of certain expectations and

the ability to draw correct inferences from them [6]. So we can use the insights of the scientific process to not fall for the false expectation that every plane bears a hidden danger. The diachronic nature of meaningful affordances [17] becomes evident when we consider how the scientific measurement only makes sense in light of past scientific insights and how it will shape our future engagement with possible observation targets.

Perception of affordances is not limited to the visual modality. The Geiger counter not only visibly displays the presence of radiation through its display but may also produce the characteristic sound. While reading a display may require knowledge about reference values and thresholds, this iconic and technologically produced sound functions as a cultural proxy for radiation danger. Technology enables us to perceive certain information previously inaccessible and integrate them in a translated form into our socio-cultural framework. This eventually relates to a form of life in the sense that:

> particular mediations by particular artefacts are part of [...] forms of life that exceed what happens at the level of the phenomenology and hermeneutics of individual use and interaction, or rather, that connects this phenomenology and hermeneutics to larger wholes and structures at the level of [...] cultures (forms of life). [7, p. 1516]

Even if only a minority of people ever had direct contact with a Geiger counter or a deadly dose of gamma radiation, they usually know, given the necessary cultural transmission, what the sound means; what to expect when hearing it [6]. Any deduced affordances are then preserved so that the knowledge about them is partially separated from directly perceiving the environmental context they emerged from.

To be fully comprehensible, the radiation measurement's visual and auditive representation must be hermeneutically translated [1]. The technologically produced stimuli must be interpreted in a "correct" way to evade the threat. This "know-how" is again transmitted culturally and distributed differently in any given population. This observation emphasizes the normative and socio-cultural dimensions of a form of life, either in terms of affordances [5],[6], technology usage [7], or particular skills and abilities [12]. As we can see, not only the ability to produce technology but also its scientific application, the transmission of knowledge gained through technology, and the normative affordance realization are deeply intertwined with their constitutive socio-cultural sphere and form of life [5],[7],[12]. Due to the cultural transmission of (technological) knowledge, human beings can spatiotemporally extend the individually acquired knowledge about affordances. This knowledge is partially obtained by the products of this sphere, i.e., skills or technologies [5],[12]. Yet, neither skills nor technologies alone create affordances. Affordances exist even if never realized or perceived by a single individual [5],[15]. Instead, they emerge from the possibility of detection and realization inside a form of life and the skills and abilities it includes as a whole [5],[12]. The whole set of available affordances constitutes the rich landscape of affordances, while the individual situational relevant possibilities of action structure this landscape into the "*field of relevant affordance*" [12]. Applying technology skillfully toward a particular object constitutes one possible relation from which the perception and eventual realization of detected affordances can emerge. At the same time, it can restructure the landscape of affordances in dependency on the technology's characteristics. The Geiger counter can neither perceive the color of the grass nor understand the ascribed value of the 79 proton element in the pot. For this device, only the gamma radiation affords to be "seen." And because this technology is part of our form of life and is attuned to our bodily capabilities, we can "see" through it and detect a highly relevant life-or-death affordance and the related expectations of a painful death.

Imagine that at some spots radiation is so low that crossing the plane is possible. In other words, a hidden and invisible maze. This maze has different affordances than the contaminated ground next to it. Even without a Geiger counter, humans can figure out the exact route by continuing to send people on the plane and eventually realizing that some places are safer than others. However, this would require an even more fatal trial-and-error process than the initial acquisition of knowledge about the plane. The use of the Geiger counter reduces possible costs. People recognize that radiation seems to fall off at the maze entrance based on the counter's display and sound. This observation alters the affordance perception of the place and directly invites action [13] at specific locations.

While the adverse effect of the radiation may be felt firsthand as an example of "natural meaning," technologically retrieved meaning, requiring correct inferences and cultural customs, may be called "non-natural" [6]. To correctly use the Geiger counter, one must perform several skillful translations, typical for the hermeneutic human-technology relationship. First, the visual display or sound must be translated into an internal danger measurement. Then this measurement must be assessed in terms of potential damage. Lastly, the affordance must be identified as relevant, evoking a state of action readiness, characterized by an organism's wish to engage in relational modulation toward their environment [12]. Note that we can identify action readiness also at the beginning of the process. The relevant affordance of the gold (spending it) in conjunction with the hidden hazard evokes the wish to change one's relation to the environment using the hermeneutic capabilities of the Geiger counter. This counter, in its instrumental realism, also performs a translation. Information about detected radiation is translated into electrical and, finally, visual and auditory information.

Let us now increase the complexity of the scenario. Assume that one has access to the Geiger counter 3000: A more sophisticated AI-powered follow-on model that can scan the entire aircraft at once and produce a detailed map with visual information about the non-hazardous path through the plane. In addition, a flashing display shows the words "Safe passage possible" and a calm voice navigates the user during the crossing. Here, the number of translations and the conducting agent differs. In the case of the ordinary Geiger counter, users must deduce certain facts themselves. They need a specific level of expertise and knowledge to arrive at the correct conclusions about the affordances of the plane. The Geiger counter 3000, however, translates the measurement of the radiation itself and directly produces a normative outcome and straightforward guide for one's actions. Using it still requires skill, but these are different from its predecessor.

Furthermore, the Geiger counter 3000 translates more information. It displays not only local information about radiation but also a spatially extended representation of the plane in front. Additionally, it gives direct navigational advice based on its perception of the affordance and in concordance with its programming. This may result in a relational shift towards an *alterity relation.* The Geiger counter 3000 is now experienced as another, equipped with certain hermeneutic capabilities that give explicit navigational and normative advice based on its perception of the world. In this way, its ability to engage with collective and shared attention is enhanced, and so is its impact on mediating expectations, relevance, and normative aspects of affordances [6]. The alterity relation differs from a simple map. Maps must be read by humans and can be more easily ignored. However, even though the Geiger counter 3000 enters the social realm of affordances and their communication, given its limited intentionality, determined by its technological structure, "its otherness remains a quasi-otherness, and its genuine usefulness still belongs to the borders of its hermeneutic capacities" [1, p. 106].

One step further to contemporary praxis is the reliance on personalized algorithms, which can also be conceived as "others." These pieces of software can profoundly shape human decision-making and give rise to individual epistemic structures that do not inevitably lead to what is most beneficial for the user [34]. Remember that correct epistemic judgments are part of a form of life [12]. So the choice of the Geiger counter 3000 and the algorithm, given an alterity relation, is evaluated in their usefulness in leading to correct epistemic conclusions in a form of life. This a posteriori evaluation, however, does not make them immune from leading to potentially catastrophic mistakes.

However, the epistemic usefulness of technology is not only a matter of the instrument and the kind of relationship but also concerns human factors. This is exemplified in a memorable but fictional scene in the TV Series "Chernobyl" [35], retelling the story of the eponymous nuclear power plant catastrophe. After the reactor accident, the technical staff tries to determine how much radiation is leaking. Their Geiger counter shows 3.6 Roentgen, which is laconically assessed as "not great, not terrible" by the chief technician in charge. However, the value of 3.6 Roentgen shown on the counter is the highest the instrument can display. A second measurement with a more potent counter reveals the radiation to be, in fact, 15,000 Roentgen. The hermeneutic relationship now has an *enigma* between technology and the world [1]. The instrument does not correctly refer to the factual world and a false affordance [27] is deduced by its interpreters. The relationship between humans and the world has become increasingly *opaque* [1]. By reading the information presented, the technical personnel is led to believe a false normative assertion, namely that there is no lethal hazard when in fact, there is. Their expectations of the conventional affordance [6] do not match the physical reality. This is where we must recognize the complex relationship between artificial hermeneutic text and the human reader: Humans rely on the information presented by the Geiger counter. They trust the device to refer to the world correctly. The assumption made by the technicians in "Chernobyl" based on the first measurement resulted in the horrible mistake of underestimating the danger by a factor of ~4000. However, it is prima facie the *correct* conclusion based on the information available to them. Although the measurement afforded to be interpreted as harmless, the technician's ability to draw the correct conclusion was limited by the materiality of the measurement instrument and its mediation of relevant information. So who is to blame in this example? The Geiger counter is neither responsible for its limiting construction nor its incorrect use.

Furthermore, it seems unfair to blame the humans who rely on the information provided by the counter. We may, of course, question their motives. The TV series does so by depicting the chief technician as incompetent and politically motivated. But this again opens up a new dimension of aspects to consider when examining conventional affordances, expectations, and human technological capabilities [6]. Especially under the suspicion of political motivation, the rich social dimension of affordances becomes evident. For this to be effective, the relevant social actors do not necessarily need to be physically present to influence engagement with situational affordances [36].

When substantially integrated into individual cognitive processes, interfering with a person's environment can have similar moral significance as interfering with them personally [37]. But does this also apply to interferences with the technology a person uses to relate to the world to access its possibilities of action? In general, the aforementioned perceived usefulness is open to deliberate manipulation. In the case of manipulation, the incorrect assessment of affordances is not rooted in the "natural" inability of the device or reader but the external exploitation of its affordance to be manipulated. If one understands the underlying technological structure and the associated hermeneutic processes, one may use the technological mediation of human perception to shape the behavior of others to their advantage. The crucial thing here is that the human body reflects the individual ecological niche to stay selectively attuned to relevant affordances [12]. Manipulating the artifacts constituting the niche potentially disrupts the coupling between the possibilities that make up the niche and the individual. Humans constructed their niche in a way that eases reasoning and problem solving [38]. Objects and technology can be used in various ways [25], but only a few possibilities have manifested themselves in actual practice. This is because this particular affordance exploitation allows for a normatively better result in a given situation [12]. However, supposing the relevant affordances are not perceivable due to deliberate manipulation, the optimal solution to a given task or problem is also not available. Like other affordances in a form of life, the utilization of such manipulation affordance depends on individual skill and the presence of other artifacts and techniques that afford technological manipulation. For an affordance to be manipulated, the affordance itself must afford manipulation. As affordances reside in the relationship between environment and individual, respectively form of life, these are the targets for any manipulation effort. The moral impact of such manipulation depends on the level of dependency and integration of the artifact, for example, in terms of using them in cognitive tasks [39]. Technologies operating phenomenological transparently and outside the range of human control or consciousness afford new ways of manipulating human behavior [40].

For exemplification, take the artwork "Google Maps Hacks" by Berlin-based artist Simon Weckert (Fig. 1). Weckert used 99 smartphones to change the Google Maps status of a street from empty to blocked by traffic jam [41]. People using Google Maps in the surrounding area were redirected to different routes in order to bypass the virtually "blocked" road. The hermeneutic relationship between the depiction of the street shown on Google Maps and the real world changed. Deduced navigational affordances about the navigability of the specific street - given that the user trusted Google Maps - were altered due to a deliberately established enigmatic relationship. The artist changed the normative aspects of the affordances that were perceived through Google Maps, proofing that a map is indeed not what it depicts [42], but also how this visualization and hermeneutic insights are vulnerable to easy manipulation due to the dynamic integration of real-time information, given knowledge of how the technology works.

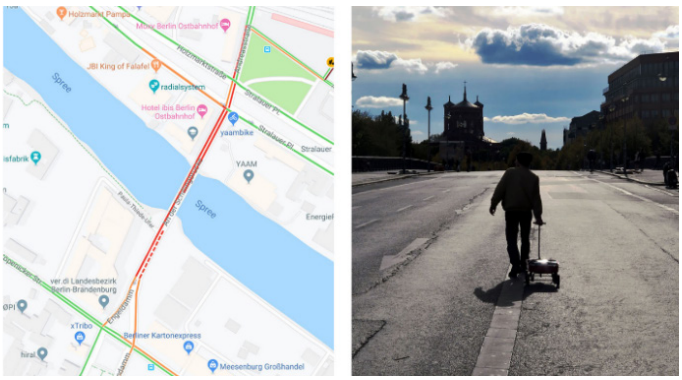


Fig. 1. "Google Maps Hacks" by Simon Weckert. Taken with permission from http://www.simonweckert.com/googlemapshacks.html

Beyond that impressive demonstration of the manipulative potential in our hermeneutic relationships, the exploitation of deduced affordances can have profound political and safety implications. One example is Global Positioning System (GPS) corruption, especially GPS spoofing [43]. All types of vehicles and services rely on GPS to navigate the environment. There are several ways in which GPS signals can be interfered with. The most frequently used type of corruption is

GPS blocking, in which the receiver's antenna is disabled or shielded. As a result, the specific antenna can no longer receive the GPS signal. Another widely used method is GPS jamming, where a different signal is transmitted with a similar frequency but higher strength, which then overlaps with the original signal. Both methods are visible to the receiver as missing GPS signals. There is also the possibility of GPS spoofing. Here, fake GPS signals are provided, indicating a different position in space and time to the user than his actual position [43]. This technology poses a severe threat for any nation or organization reliant on GPS, concentrating (geo-) political power in the hands of the political players capable of GPS corruption.

Prima facie, the misdirection of a vehicle seems to be neglectable. But consider the possible impact if the vessel in question transports military or humanitarian supplies or a political VIP. This example demonstrates that: "navigation systems are indeed instruments for realizing one's intentions and goals, [how] they also embody moral values like safety, transform the experience of our environment, and have unintended consequences on our onboard cognitive capabilities." [39, p. 26]. The moral value of the navigational system is constituted in relation to its affordances [15],[16]. Thus, the ethical severity of any manipulation depends on the same value. Although only affecting a single device, the technological effect of corrupting a single GPS can spread into the (global) sociopolitical sphere and its organizational structures, whose regularities and principles are different and more challenging to predict than in the immediate context of technology [8]. This scenario emphasizes the need for a thorough examination of technology that encompasses not only its primary direct effect on the perception of affordances but also the secondary effects on the techno-political dimensions of the form of life to which those affordances relate. Here some consequences may not be instantaneously apparent. Furthermore, it shows how instrumental realism based on a particular technology can be outwitted by the use of another technology, adding complexity to the human-technology(-technology) relationship. Consequently, technologically sophisticated political players can force other agents into certain types of behavior, not by direct threat or negotiation, but by simply deceiving them about the range of possibilities of action in a given scenario.

Through the conscious and technological establishment of an enigmatic hermeneutic relationship, technologies can not only reveal but also be used to deceive users about the affordances of their environment. This is made feasible by the fact that the affordances themselves can be manipulated, which in turn is only possible because the technologies with which we enter into a hermeneutic relationship are part of a particular form of life and because users expect the device to relate correctly to the environment. Without a GPS device or insufficient skills of using it, one could not be deceived about the affordances as the relevant relation would not be present. Moreover, if we expect the device not to function properly, users will not trust the information it provides to them. By deducing a certain possibility or non-possibility of action as well as the correct way of executing it, people will engage in different patterns of behavior in relation to their socio-material environment and form of life. If picked up by a sufficient amount of people, this engagement will manifest itself in a new practice or will alter an already existing one. Thus, manipulating the perception of affordances using technology can directly influence which affordances are available within a human form of life. The action derived from these interrelated affordances can then change the human ecological niche, i.e., how humans live.

IV. Enhancing the Perspective

It should be clear now how the perception and realization of affordances can be mediated technologically and even manipulated. From there, the question arises of what happens when technological manipulation's power is directed toward humans themselves. Affordances depend on the environment but also the abilities of specific individuals [5],[6],[12]. So, what when these abilities are changed technologically? This touches on the concept of *Human Enhancement;* the effort to create "better" humans through the application of technology [9],[44]. Definitions of Human Enhancement often center around creating new capabilities and capacities through science and technology [45]–[47]. If enhancements and these new capabilities will inevitably lead to "better" humans is subject of controversial debate [48]–[51]. Independent of the claim about its final normative end state, we may examine the concept of Human Enhancement through the lens of affordances. Or, more precisely, how the implementation of technology is connected to the possibilities of action the enhanced person may perceive and realize.

Which capabilities or features of humans are up to enhancement often remains vague. Yet, it appears obvious that improving the capabilities of mind and body goes hand in hand with the emergence of new affordances since they heavily depend on the relationship between the environment, the individual bodily characteristics, and the set of skills [12]. Alternation of one's body is one of the most straightforward ways to change individual affordances. The most infamous means of transformation discussed in the Human Enhancement debate are the so-called NBIC (Nanotechnology, biotechnology, and cognitive science) methods. These enhancement means are rejected or endorsed primarily due to their hypothetical transformative power and consequences on the human condition [51],[52]. The transformative potential of Human Enhancement can be expressed in terms of affordances: as an extension of the human opportunities of action through deliberate and direct manipulation of the human body and mind.

Returning to the example of radiation: there are ongoing discussions about Human Enhancement during space missions to improve resistance against cosmic radiation [53]. This is one of many examples of how technological interventions can adapt the human body to new environments that offer new possibilities of action. This is the opposite of the typical evolutionary strategy of adapting the environment [54]. Instead of the environment, we are adapting ourselves [55].

The debate about Human Enhancement often revolves around whether the technological intervention must be implemented within the human body or can remain external [45],[56]. Regardless of its necessity for definition, the internal implementation of a technological device may realize a so-called *cyborg relation,* which brings forth a new entity equipped with a new *hybrid intentionality* composed of the features of the human and the machine [21]. It is worth noting that the term "cyborg" originally described an individual that used exogenous components to adapt to new and potentially hazardous environments [57], altering the environment's affordance to support survival.

Cyborg relation [21]*:* (*Human*/*Technology*) → *World*

An example of this can be found among the community of "Biohackers," who actively share and promote knowledge about how to self-implant small magnets to gain magnetoreception [58]. Gaining new senses is a clear example of Human Enhancement and allows for detecting and realizing new affordances. These new affordances not only emerge because the technological improvement of human capabilities may enable the respective individual to execute different behavior but also because these special technologies have become part of a certain form of life, enriching it with new skills and abilities. Furthermore, merging technology and human is only possible because the biological tissue and related perceptual processes can be attuned to incorporate other materials and sensory inputs. Technological enhancement is thus related to the possibilities for action that the

environment offers the individual and their constitution, as well as the embedding of that same individual in a larger socio-material structure.

Some argue that the next step in the evolution of *Homo sapiens* will be its merging with technology [59]. Others believe that humans were, in some sense, cyborgs all along [60] or that the aspiration of merging with the machine and transforming human capabilities fulfills a romantic desire [61]. Undoubtedly, the human body and mind themself afford to be changed. Tools afford to be integrated into the human body schema [60],[62], a process moderated by the expertise of the user [63]. Hence, besides the literal embodiment, realized by the spatial implementation of a technology, there are other dimensions of embodiment, depending on the motoric and affective attitude towards the device [64]. Here, the degree of embodiment directly influences action awareness and planning [64] and, therefore, the perception and realization of affordances. Consider, for example, a robotic third thump. Albeit external, it psychologically merges with the body representation of the user [65], who is now able to pick up new affordances from the environment. It is easy to see how the wide dissemination of such a technology would change how humans engage with their environment.

Thinking of humans as "natural born cyborgs" [60], we may even conclude that a considerable proportion of the way humans interact with their environment and eventually reconstruct it to meet their demands is due to the psychological and physiological embodiment of things and the embedding of the human mind and activity into the technological sphere. Adopting the perspective of humans as "profoundly embodied agents," constantly renegotiating the boundaries between environment and body [66], opens up a new and enhanced perspective on affordances. It does not only highlight how affordances emerge from the relationship between (technologized) body and environment but also the transformative nature of technology: Not only do humans perceive and realize affordances *through* technology, but they are granted new ones *by* technology.

V. Affordance in the Technological Form of Life

We provided several examples to demonstrate the transformative power of technology on human perception and action. As stated before, human activity is embedded in and exercised by technology, and it is this (perceptual) activity that is crucial for the perception of affordances in general [2],[24]. This is of particular concern when the transformative power of technologies is directed against the life forms that have constructed them in the first place. Here, the technology not only mediates human perception and action but also directly interferes with its user's psycho-physiological constitution. So-called Human Enhancement technologies are, in some way, already widely used [8],[67],[68] and, therefore, part of the human technological form of life.

Using technology, human beings commit to a specific form of life, comprising stable patterns of behavior [7], which then stands in relation to the possibilities of action afforded to it [5]. Overall affordance modulation research has to consider various intra- and interpersonal contextual factors in relation to the presented task [69]. It has been argued that there are at least two ways of changing the available affordances to an organism: Changing the material environment or altering its form of life or set of abilities [6]. Given the transformative power of technology, both on the level of perception and Human Enhancement, we must add *technological mediation* and *technological alteration of the body* inside a form of life to that list.

Comprehensively and conceptually, technology can bring forth affordances hitherto not perceivable to human beings. Ontologically speaking, these affordances start to exist once they are, in principle, detectable by the skills and abilities of an individual or the general capacities of a form of life [5]. However, the possibility for action does not instantaneously make these affordances relevant to the other individuals engaged in the same form of life. To translate a possibility from the relevant field of affordances to the rich landscape, people must engage in communicative behavior and teaching about the meaning of a particular affordance in a given situation. The remarkable fact of the human form of life is that it only needs one person with the adequate scientific instrument to inform other species members about the hidden danger of a deceptively empty plane. However, it also takes one person equipped with 99 smartphones to fool an entire online community about the affordances of one street, rendering the driving affordance of the same street irrelevant for navigational purposes.

Using technology, humans are changing their relationship with the world and are introducing a new entity in the reciprocal dynamic of this dyade. Technology in the form of a concrete artifact constitutes a new tripartite relationship between this artifact, its environment, and the user and his psychological and physiological characteristics [24]. Given an alterity relation, the artifact becomes a quasi-other and a social proxy in the already socio-cultural sphere of affordance perception. In a more intimate cyborg relationship, humans and technology merge. This process of cyborgization is accompanied by moral concerns and ethical challenges [70]. Some fear that enhancing human capabilities through technological means may even lead to a state of "hyperagency" in which enhancement provides too many opportunities to manipulate internal affairs, which negatively affects how we interact with the world and should therefore be constrained [71]. Seen from the perspective of affordances as connected to the value of a technology [15], some negative attitudes toward Human Enhancement may be due to the possibilities of action the enhancement may provide in a specific context.

Given the necessary expertise and knowledge, every side of the aforementioned tripartite relationship of artifact, human, and environment is susceptible to manipulations. Technological manipulation may lead to the creation of *deceptive affordances*, meaning the conscious misdirection about possibilities of action in a particular setting. Technology can both reveal and veil affordances. This general amplifying and reductive effect of technology may even occur simultaneously and unpreventably [1]. The revealing power of a technology can exceed its veiling effect. However, it is essential to remember this co-dependency when assessing hermeneutic technological relationships and the perception and actions that emerge from them.

Overall, "[t]he impact of technological mediation, [...] results not only from the roles human beings allow technologies to play in their lives but also from the characteristics of technologies that help to shape their mediating roles" [14, p. 89]. Deliberately designed technological mediations due to the actions they may elicit are, therefore, of severe moral concern [14]. Depending on the particular technology, technological mediation could already be part of a specific form of life, interwoven in the regular ways of doing things and engaging with technology on a more general level. Interestingly this can go so far that humans are not even aware of the mediation [13]. If done effectively and sustained, the manipulation of technological mediation can change not only a particular behavior but also a form of life [7]. Accordingly, to understand any ongoing change, we must examine the use of the specific technology in its form of life and thus the affordances it offers to this form and the dynamic changes within this relationship. This is where we must consider a variety of socio-technological forces and be aware of the very nature of a successful manipulation, the unawareness of it happening. The more we rely on technology and its hermeneutic qualities and the enhancement of our capabilities, the more severe the possible damage of any affordance manipulation.

However, in an open debate about the ethical implications of the normative nature of affordances and the technological mediation of their perception, one must understand that withdrawing from the pervasive influence of technology is impossible. Any regulation of technology happens "from within" not only in terms of the actual mediation process but also in a specific form of life [7],[14]. Drawing on the metaphor of a form of life as a river bed [7] and the role affordances may play in assessing the value of a technology [15],[16], it is this form of life that brings forth the technology in question in the first place. By using certain possibilities of action, the material environment is manipulated to create new technology. This technology then influences the perception and realization of present and hidden affordances. While doing so, the affordances of the particular technology in a given context may constitute its value [15], thus prompting eventual regulation based on this value. Therefore, potentially disruptive technologies establish a new normativity or influence existing ones [25]. In other words, by connecting affordances to the concept of the (technological) form of life, scholars can shine new light on how technologies alter perception and action. It opens up a new perspective on how humans regularly do things and how the things humans regularly do recursively influence how humans (will) live.

This is not only an abstract philosophical issue but also concerns the political implication of technological mediation of the perception of affordances. Here, the recourse to the framework of postphenomenology is suitable as it "may not have an inherent *politics*, but it certainly is *political* in that it paves the way for phenomenologically informed interventions" [18, p. 530, italics from the original]. We must, therefore, further examine how technological mediation influences our form of life, the related affordances, and, thus, the whole spectrum of how we engage with the world normatively and perceptually. Even if an artifact is not intentionally designed for manipulation purposes, the affordances of an artifact can make up its moral value [15],[16],[25]. We must face how human behavior is interwoven with the material environment to which humans uphold a recursive relationship. That is, humans use technologies to gain new insights about themselves but also the same environment. A side effect of this improved understanding is not only progress in the romantic quest for positive epistemic knowledge but also insights into how we can influence the behavior of others.

Moreover, we learn new ways of manipulating ourselves. Human Enhancement technologies afford to change aspects of the human body and mind. The ethical debate about these technologies [37], [38], [55]–[57] can benefit from a point of view rooted in an affordance approach. Considering that an ecological niche can be conceived as a set of affordances in a specific environment to a particular time [6] and that the human environment is accumulated with technology of all scopes, kinds, and varying complexities [1],[8],[72], we conclude that technologically induced alteration through Human Enhancement is already part of the human niche. Altering ourselves with technology is part of how we live, and in light of the aforementioned ethical debate about the benefits or drawbacks of Human Enhancement in general, it also addresses the question Verbeek [13] posed for technologies in general: "How to live?"

VI. Conclusion

The human form of life is technological, and technology can change it. By linking the concept of affordances to the mediating perspective of postphenomenology, we can reveal crucial aspects in our understanding of how the perception of affordances functions in a technological context and eventually influences the human form of life and ecological niche. Moreover, new affordances can also arise when humans alter themselves by means of technology.

We have emphasized not only the role of the used technology but also the need for compatibility with the characteristics of its human user and the hermeneutic act of "reading" the provided information. Going one step further, the merging of humans and technology, constituting a new entity, is accompanied by new affordances. Whether the process of cyborgization will eventually lead to a form of sophisticated cyborg-life remains an open question. Considering our heavy reliance on technology in virtually every aspect of our lives, technology affords various ways of manipulation. Not only in terms of manipulating the environment or oneself, but also by deceiving humans about what they can do in a given situation.

In the original conception of Spider-Man, Peter Parker obtained the ability to shoot his webs not through the bite of the radioactive spider but rather through self-build, wrist-attached "web-shooters." This enhancing cyborg relationship between technology and teenage boy suddenly brought new affordances to the mind of Peter Parker. Skyscrapers now afforded the attachment of spider webs, and street canyons afforded swinging. This concluding anecdote is meant not only to exemplify the perception and action transformative power of human enhancement technologies but to serve as a reminder that with great power comes great responsibility.

Acknowledgments

The authors want to thank Albrecht Kleinlein for his proofreading. Special thanks to Claus-Christian Carbon for his guidance in the creation process of the manuscript.

This is an extended version of a text, that was presented at the 1st International Conference on Disruptive Technologies Tech Ethics and Artificial Intelligence (DiTTEt 2021), held from 9/15-9/17 in Salamanca (Spain) and can be found here: DOI: 10.1007/978-3-030-87687-6_15. Modifications include improving language and style, elaborating arguments, and adding the section discussing Human Enhancement.

References

[1] Ihde, D., *Technology and the lifeworld. From garden to earth,* Indiana University Press, Bloomington, 1 Jan. 1990, 226.

[2] Gibson, J. J., *The theory of affordances,* Houghton Mifflin, Boston, 1 Jan. 1979.

[3] Withagen, R., and Costall, A., "What does the concept of affordances afford?," *Adaptive Behavior*, 1 Jan. 2021. doi: 10.1177/1059712320982683.

[4] Lanamäki, A., Devinder, T., and Stendal, K., "What does a chair afford? A Heideggerian perspective of affordance," *Selected Papers of the IRIS, Issue Nr 6*, 1 Jan. 2015, URL: https://aisel.aisnet.org/iris2015/2.

[5] Rietveld, E., and Kiverstein, J., "A rich landscape of affordances," *Ecological Psychology*, Vol. 26, No. 4, 1 Jan. 2014, pp. 325–352. doi: 10.1080/10407413.2014.958035.

[6] Ramstead, M. J. D., Veissière, S. P. L., and Kirmayer, L. J., "Cultural affordances: Scaffolding local worlds through shared intentionality and regimes of attention," *Frontiers in Psychology*, Vol. 7, 1 Jan. 2016. doi: 10.3389/fpsyg.2016.01090.

[7] Coeckelbergh, M., "Technology games: Using Wittgenstein for understanding and evaluating technology," *Science and Engineering Ethics*, Vol. 24, No. 5, 1 Jan. 2018, pp. 1503–1519. doi: 10.1007/s11948-017-9953-8.

[8] Allenby, B. R., and Sarewitz, D. R., *The techno-human condition,* MIT Press, Cambridge, Mass., 1 Jan. 2011, 222.

[9] Coeckelbergh, M., *Human being @ risk. Enhancement, technology, and the evaluation of vulnerability transformations,* Springer, 1 Jan. 2013.

[10] Ihde, D., and Malafouris, L., "Homo faber revisited: Postphenomenology and Material Engagement Theory," *Philosophy & Technology*, Vol. 32, No. 2, 1 Jan. 2019, pp. 195–214. doi: 10.1007/s13347-018-0321-7.

[11] Malafouris, L., *How things shape the mind. A theory of material engagement,* The MIT Press, Cambridge, Massachusetts, 1 Jan. 2013, 304.

[12] Rietveld, E., Denys, D., and van Westen, M., "Ecological-enactive cognition as engaging with a field of relevant affordances: The skilled intentionality framework (SIF)," *The Oxford Handbook of 4E Cognition*, edited by A. Newen, L. de Bruin and S. Gallagher, Oxford University Press, 1 Jan. 2018, pp. 41–70.

[13] Verbeek, P.-P., *What things do*, Pennsylvania State University Press, University Park, PA, 1 Jan. 2005.

[14] Verbeek, P.-P., *Moralizing technology. Understanding and designing the morality of things*, University of Chicago Press, 1 Jan. 2011, 200.

[15] Klenk, M., "How do technological artefacts embody moral values?," *Philosophy & Technology*, Vol. 34, No. 3, 1 Jan. 2021, pp. 525–544. doi: 10.1007/s13347-020-00401-y.

[16] Tollon, F., "Artifacts and affordances: from designed properties to possibilities for action," *AI & SOCIETY*, 1 Jan. 2021. doi: 10.1007/s00146-021-01155-7.

[17] Dings, R., "Meaningful affordances," *Synthese*, Vol. 199, 1-2, 1 Jan. 2021, pp. 1855–1875. doi: 10.1007/s11229-020-02864-0.

[18] Aagaard, J., "Introducing postphenomenological research: a brief and selective sketch of phenomenological research methods," *International Journal of Qualitative Studies in Education*, Vol. 30, No. 6, 1 Jan. 2016, pp. 519–533. doi: 10.1080/09518398.2016.1263884.

[19] Ihde, D., *Embodied technics*, Automatic Press, 1 Jan. 2010.

[20] Ihde, D., *Technics and praxis*, D. Reidel, Dodrecht, 1 Jan. 1979.

[21] Verbeek, P.-P., "Cyborg intentionality: Rethinking the phenomenology of human–technology relations," *Phenomenology and the Cognitive Sciences*, Vol. 7, No. 3, 1 Jan. 2008, pp. 387–395. doi: 10.1007/s11097-008-9099-x.

[22] Verbeek, P.-P., "Toward a theory of technological mediation: A program for postphenomenological research," *Technoscience and postphenomenology. The manhattan papers*, edited by J. K. Berg, O. Friis and R. C. Crease, Lexington Books, Lanham, 1 Jan. 2016.

[23] Chemero, A., "An outline of a theory of affordances," *Ecological Psychology*, Vol. 15, No. 2, 1 Jan. 2003, pp. 181–195. doi: 10.1207/S15326969ECO1502_5.

[24] Stoffregen, T. A., and Mantel, B., "Exploratory movement and affordances in design," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, Vol. 29, No. 3, 1 Jan. 2015, pp. 257–265. doi: 10.1017/S0890060415000190.

[25] Boer, B. de, "Explaining multistability: postphenomenology and affordances of technologies," *AI & SOCIETY*, 1 Jan. 2021. doi: 10.1007/s00146-021-01272-3.

[26] Withagen, R., Poel, H. J. de, Araújo, D., and Pepping, G.-J., "Affordances can invite behavior: Reconsidering the relationship between affordances and agency," *New Ideas in Psychology*, Vol. 30, No. 2, 1 Jan. 2012, pp. 250–258. doi: 10.1016/j.newideapsych.2011.12.003.

[27] Gaver, W. W., "Technology affordances," *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91*, edited by S. P. Robertson, G. M. Olson and J. S. Olson, ACM Press, New York, New York, USA, 1 Jan. 1991, pp. 79–84.

[28] Heras-Escribano, M., and Pinedo, M. de, "Are affordances normative?," *Phenomenology and the Cognitive Sciences*, Vol. 15, No. 4, 1 Jan. 2016, pp. 565–589. doi: 10.1007/s11097-015-9440-0.

[29] Gier, N. F., "Wittgenstein and forms of life," *Philosophy of the Social Science*, Vol. 10, 1 Jan. 1980, pp. 241–258.

[30] Froese, T., "Scientific observation is socio-materially augmented perception: Toward a participatory realism," *Philosophies*, Vol. 7, No. 2, 1 Jan. 2022, p. 37. doi: 10.3390/philosophies7020037.

[31] Warren, W. H., "Perceiving affordances: visual guidance of stair climbing," *Journal of experimental psychology. Human perception and performance*, Vol. 10, No. 5, 1 Jan. 1984, pp. 683–703. doi: 10.1037//0096-1523.10.5.683.

[32] Ihde, D., "Stretching the in-between: Embodiment and beyond," *Foundations of Science*, Vol. 16, 2-3, 1 Jan. 2011, pp. 109–118. doi: 10.1007/s10699-010-9187-6.

[33] Boer, B. de, Molder, H. t., and Verbeek, P.-P., "'Braining' psychiatry: an investigation into how complexity is managed in the practice of neuropsychiatric research," *BioSocieties*, 1 Jan. 2021. doi: 10.1057/s41292-021-00242-8.

[34] Heersmink, R., "Varieties of artifacts: Embodied, perceptual, cognitive, and affective," *Topics in Cognitive Science*, Vol. 13, No. 4, 1 Jan. 2021, pp. 573–596. doi: 10.1111/tops.12549.

[35] Renk, J., *Chernobyl* [TV-series], 1 Jan. 2019.

[36] Rietveld, E., and Brouwers, A. A., "Optimal grip on affordances in architectural design practices: an ethnography," *Phenomenology and the Cognitive Sciences*, Vol. 16, No. 3, 1 Jan. 2017, pp. 545–564. doi: 10.1007/s11097-016-9475-x.

[37] Clark, A., and Chalmers, D. J., "The extended mind," *Analysis*, Vol. 58, No. 1, 1 Jan. 1998, pp. 7–19.

[38] Wheeler, M., and Clark, A., "Culture, embodiment and genes: unravelling the triple helix," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, Vol. 363, No. 1509, 1 Jan. 2008, pp. 3563–3575. doi: 10.1098/rstb.2008.0135.

[39] Heersmink, R., "Extended mind and cognitive enhancement: moral aspects of cognitive artifacts," *Phenomenology and the Cognitive Sciences*, Vol. 16, No. 1, 1 Jan. 2017, pp. 17–32. doi: 10.1007/s11097-015-9448-5.

[40] Wheeler, M., "The reappearing tool: transparency, smart technology, and the extended mind," *AI & SOCIETY*, Vol. 34, No. 4, 1 Jan. 2019, pp. 857–866. doi: 10.1007/s00146-018-0824-x.

[41] Weckert, S., "Google Maps hacks," URL: http://www.simonweckert.com/googlemapshacks.html. Last accessed 11/01/2022 [retrieved 11 January 2022].

[42] Korzybski, A., *Science and sanity; an introduction to Non-Aristotelian systems and general semantics*, Lancaster, 1 Jan. 1933.

[43] Warner, J. S., and Johnston, R. G., "GPS spoofing countermeasures," *Homeland Security Journal*, Vol. 25, No. 2, 1 Jan. 2003, pp. 19–27, URL: https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-03-6163.

[44] Coeckelbergh, M., "Human development or human enhancement? A methodological reflection on capabilities and the evaluation of information technologies," *Ethics and Information Technology*, Vol. 13, No. 2, 1 Jan. 2011, pp. 81–92. doi: 10.1007/s10676-010-9231-9.

[45] Allhoff, F., Lin, P., Moor, J., and Weckert, J., "Ethics of human enhancement: 25 Questions & answers," *Studies in Ethics, Law, and Technology*, Vol. 4, No. 1, 1 Jan. 2010, pp. 1–39. doi: 10.2202/1941-6008.1110.

[46] James, D., "The ethics of using engineering to enhance athletic performance," *Procedia Engineering*, Vol. 2, No. 2, 1 Jan. 2010, pp. 3405–3410. doi: 10.1016/j.proeng.2010.04.165.

[47] Buchanan, A. E., *Beyond humanity? The ethics of biomedical enhancement*, Oxford Univ. Press, Oxford, 1 Jan. 2011, 286.

[48] Browne, T. K., and Clarke, S., "Bioconservatism, bioenhancement and backfiring," *Journal of moral education*, Vol. 49, No. 2, 1 Jan. 2020, pp. 241–256. doi: 10.1080/03057240.2019.1576125.

[49] Bostrom, N., "Transhumanist values," *Ethical Issues for the 21st Century*, edited by F. Adams, Philosophical Documentation Center Press, 1 Jan. 2003, pp. 3–14.

[50] Hauskeller, M., *Better humans? Understanding the enhancement project*, Acumen, Durham, 1 Jan. 2013, 223.

[51] Dupuy, J.-P., "Cybernetics Is antihumanism: Advanced technologies and the rebellion against the human condition," *H± transhumanism and its critics*, edited by G. R. Hansell and W. Grassie, Xlibris, Philidelphia, 1 Jan. 2011, pp. 227–248.

[52] Sandberg, A., "Morphological Freedom - Why we not just want It, but need it," *The Transhumanist Reader*, edited by M. More and N. Vita-More, John Wiley & Sons, Chichester, West Sussex, 1 Jan. 2013, pp. 58–64.

[53] Szocik, K., Norman, Z., and Reiss, M. J., "Ethical challenges in human space missions: A space refuge, scientific value, and human gene editing for space," *Science and Engineering Ethics*, Vol. 26, No. 3, 1 Jan. 2020, pp. 1209–1227. doi: 10.1007/s11948-019-00131-1.

[54] Kirsh, D., "Adapting the environment instead of oneself," *Adaptive Behavior*, Vol. 4, No. 3/4, 1 Jan. 1996, pp. 415–452. doi: 10.1177/105971239600400307.

[55] Pustovrh, T., Mali, F., and Arnaldi, S., "Are better workers also better humans? On pharmacological cognitive enhancement in the workplace and conflicting societal domains," *NanoEthics*, Vol. 12, No. 3, 1 Jan. 2018, pp. 301–313. doi: 10.1007/s11569-018-0332-y.

[56] Bostrom, N., and Roache, R., "Ethical issues in human enhancement," *New waves in applied ethics*, edited by J. Ryberg, T. Petersen and C. Wolf, Pelgrave Macmillan, 1 Jan. 2008, pp. 120–152.

[57] Clynes, M. E., and Kline, N. S., "Cyborgs and space," *Astronautics*, 1 Jan. 1960, 26–27/74-76.

[58] Yetisen, A. K., "Biohacking," *Trends in biotechnology*, Vol. 36, No. 8, 1 Jan. 2018, pp. 744–747. doi: 10.1016/j.tibtech.2018.02.011.

[59] Barfield, W., "The process of evolution, human enhancement technology,

and cyborgs," *Philosophies*, Vol. 4, No. 1, 1 Jan. 2019, pp. 1–14. doi: 10.3390/philosophies4010010.

[60] Clark, A., *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*, Oxford University Press, Oxford, 1 Jan. 2003.

[61] Coeckelbergh, M., *New romantic cyborgs. Romanticism, information technology, and the end of the machine*, MIT Press, Cambridge, 1 Jan. 2017, 332.

[62] Martel, M., Cardinali, L., Roy, A. C., and Farnè, A., "Tool-use: An open window into body representation and its plasticity," *Cognitive Neuropsychology*, Vol. 33, 1-2, 1 Jan. 2016, pp. 82–101. doi: 10.1080/02643294.2016.1167678.

[63] Weser, V. U., and Proffitt, D. R., "Expertise in tool use promotes tool embodiment," *Topics in Cognitive Science*, Vol. 13, No. 4, 1 Jan. 2021, pp. 597–609. doi: 10.1111/tops.12538.

[64] Vignemont, F. de, "Embodiment, ownership and disownership," *Consciousness and Cognition*, Vol. 20, No. 1, 1 Jan. 2011, pp. 82–93. doi: 10.1016/j.concog.2010.09.004.

[65] Kieliba, P., Clode, D., Maimon-Mor, R. O., and Makin, T. R., "Robotic hand augmentation drives changes in neural body representation," *Science robotics*, Vol. 6, No. 54, 1 Jan. 2021. doi: 10.1126/scirobotics.abd7935.

[66] Clark, A., "Re-inventing ourselves: The plasticity of embodiment, sensing, and mind," *Journal of Medicine and Philosophy*, Vol. 32, 1 Jan. 2007, pp. 263–282. doi: 10.1080/03605310701397024.

[67] Döbler, N. A., and Carbon, C.-C., "Vaccination against SARS-CoV-2: A human enhancement story," *Translational Medicine Communications*, Vol. 6, 1 Jan. 2021. doi: 10.1186/s41231-021-00104-2.

[68] Greely, H. T., "Regulating human biological enhancements: Questionable justifications and international complications," *Santa Clara Journal of International Law*, Vol. 4, No. 2, 1 Jan. 2006, pp. 87–110.

[69] Carbon, C.-C., "Psychology of design," *Design Science*, Vol. 5, No. 26, 1 Jan. 2019. doi: 10.1017/dsj.2019.25.

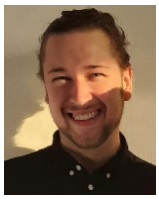
[70] Ireni-Saban, L., and Sherman, M., "Cyborg ethics and regulation: ethical issues of human enhancement," *Science and Public Policy*, 1 Jan. 2021. doi: 10.1093/scipol/scab058.

[71] Danaher, J., "Hyperagency and the good life – Does extreme enhancement threaten meaning?," *Neuroethics*, Vol. 7, No. 2, 1 Jan. 2014, pp. 227–242. doi: 10.1007/s12152-013-9200-1.

[72] Haff, P., "Humans and technology in the Anthropocene: Six rules," *The Anthropocene Review*, Vol. 1, No. 2, 1 Jan. 2014, pp. 126–136. doi: 10.1177/2053019614530575.

Niklas A. Döbler

Niklas A. Döbler holds a Master's and Bachelor's degree in psychology from the University of Bamberg. He currently works on his thesis about human enhancement, Transhumanism, and Bioengineering and is supervised by Claus-Christian Carbon. His further research interests include the psychological aspects of the Search for Extraterrestrial Intelligence (SETI), space psychology, and human-technology interaction. In his work, he strives to combine both, empirical and theoretical insights from various disciplines. orcid.org/0000-0001-7935-727X



Clemens Bartnik

Clemens obtained his Master in Psychology (M.Sc.) working towards quantitative comparison of visual saliency maps of convolutional neural networks and those of human beings, where he was supervised by Ute Schmid (University of Bamberg). And is now a second year PhD candidate in the Video & Image Sense Lab at the Informatics Institute of the University of Amsterdam, under the supervision of Iris Groen and Cees Snoek. His research focuses on leveraging computational modeling and neuroimaging techniques to understand representations of navigational affordances in the human visual system and computational models. He is inspired by the idea to combine the successes of state-of-the-art computer vision algorithms with classical approaches of measuring human behavior and neuroimaging to better understand how we perceive the world around us.

A Model for Planning TELCO Work-Field Activities Enabled by Genetic and Ant Colony Algorithms

João Henriques, Filipe Caldeira *

Informatics Department, Polytechnic of Viseu, 3504-510 Viseu (Portugal)
Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra (Portugal)
CISeD—Research Centre in Digital Services, Polytechnic of Viseu, 3504-510 Viseu (Portugal)

Received 5 April 2022 | Accepted 4 July 2022 | Early Access 19 August 2021

Abstract

Telecommunication Company's (TELCO) are continuously delivering their efforts on the effectiveness of their daily work. Planning the activities for their workers is a crucial sensitive, and time-consuming task usually taken by experts. This plan aims to find an optimized solution maximizing the number of activities assigned to workers and minimizing the inherent costs (e.g., labor from workers, fuel, and other transportation costs). This paper proposes a model that allows computing a maximized plan for the activities assigned to their workers, allowing to alleviate the burden of the existing experts, even if supported by software implementing rule-based heuristic models. The proposed model is inspired by nature and relies on two stages supported by Genetic and Ant Colony evolutionary algorithms. At the first stage, a Genetic Algorithms (GA) identifies the optimal set of activities to be assigned to workers as the way to maximize the revenues. At a second step, an Ant Colony algorithm searches for an efficient path among the activities to minimize the costs. The conducted experimental work validates the effectiveness of the proposed model in the optimization of the planning TELCO work-field activities in comparison to a rule-based heuristic model.

Keywords

Ant Colony, Genetic Algorithms, Route Optimization, TELCO.

DOI: 10.9781/ijimai.2022.08.011

I. Introduction

The TELCO are putting significant efforts into the optimization of their current operations in order to strengthen business competitiveness. One key aspect that should be addressed is optimizing the activities assigned to workers to be executed at remote locations. Preparing an optimized work plan is a challenging and complex task, even for specialists, maximizing the number of activities to be assigned to workers while keeping low as possible the use of resources, including labor, fuel, and vehicles. Despite this, computing an optimized plan is a time-consuming task taken by experts in a time-consuming iterative trial and error process, even if supported by software implementing rule-based heuristic models.

Workers are usually assigned to geographical areas, departing from their base stations to execute the planned activities within an expected period. Typically the base location is the location they return to at the end of the day. In order to assign activities to a worker, the skills required by the activities and the worker skills should match.

Thus, the availability of work at different locations is a crucial resource that organizations should carefully manage. They set up plans to optimize and maximize the number of activities for every workday. A solution must consider the time to run the foreseen activities, the duration and kilometers of the journey to and from each activity location, and the labor, fuel, and kilometers in maintenance. The quality of a candidate solution depends on the number of activities as revenue, while costs result from labor and the distance between different locations. Thus, a solution composed of revenue and cost means that it has a monetary value. Thus, it will be possible to compare the different solutions from other models despite their different nature and structure.

This work takes inspiration from nature to propose an optimization model for planning the activities of TELCO. The model incorporates heuristics as key knowledge retrieved from businesses to schedule workers' activities. The application of Evolutionary Algorithms (EA) is explored as an alternative to the common use of rule-based heuristic models supported in two stages. In the first stage, the best set of activities assigned to workers supported by the use of GA is selected. The second stage defines the order by which each activity should be executed according to their different locations by optimizing the distance the workers should run, supported by the use of Ant Colony Optimization (ACO). This optimization requires computing the best route to the different activities locations.

Beyond this section, section II presents the background and the key concepts. Section III presents the related work. Section IV, describes the implemented model. Section V presents the experimental work. Section VI discusses the achieved results. Section VII concludes the paper.

II. Background

This section provides the background on the adopted methodology to optimize the selection of the activities and the order to execute them. For that purpose, it is explored the use of Genetic Algorithms and Ant Colony algorithms.

* Corresponding author.
E-mail addresses: joaohenriques@estgv.ipv.pt (J. Henriques), caldeira@estgv.ipv.pt (F. Caldeira).

A. Genetic Algorithm

The use of GA represents an alternative to the rule-based model by exploring the space of solution containing the optimized set of activities to be assigned to workers. Such activities result from the scheduling process that minimizes the duration of their activities.

The GA takes individuals containing the encoding solutions in chromosomes by evaluating the fitness function to create a new offspring supported by the crossover operator [1], [2]. The operation performs a random mutation when creating a new offspring. The chromosomes are commonly encoded as strings containing binary, real-valued, integer, octal, or hexadecimal numbers. Initially, a random population is created, providing an ample search space for potential solutions.

A fitness function scores the individuals in the current population, determining their survival for the next generation. The individuals returning the higher scores are the ones who are likely to be selected for the next generation. The individual score depends on how well chromosomes can solve the problem at hand, mostly done using probabilistic methods supported by evolutionary computing research, such as *roulette wheel, rank selection* and *tournament.*

B. Ant Colony Optimization Algorithm

Ant species can find the shortest path between a food source and the nest. The ACO simulates the behavior of ants as agents collectively searching the space for food while sharing information among them to achieve reasonable solutions [3]. Because they drop pheromones every time they bring food, shorter paths are more likely to have more significant amounts of pheromones, hence optimizing the solution. Ants select the next location to follow, depending on the distance and the amount of pheromone in the path. Ants have some properties like memory and sight. Their memory helps them to, among others, save the locations they visited, the distance they traveled, and the shortest path they saw. Sight allows them to know the possible end locations j and the distance d_{ij} to travel from their location at a given point i, with $j \neq i$. Ants are forced to make complete tours by maintaining the information about the previously visited locations in a tabu search [4].

Traveling Salesman Problem (TSP) aims to compute the optimized path considering the locations to visit. Due to this complex combination problem, a meta-heuristic approach optimization is used to find the shortest route from the nest to the food source.

The equation 2 provides the mathematical representation of the TSP problem, where n is the number of cities. The optimal solution of π with index nodes 1, 2, ..., n, such as the length of π is minimal and d is the distance between those nodes index, The $\Pi\{1, 2, ..., n\}$ indicates all the permutations 1, 2, ..., n [1]

$$\text{minimizef}\,(\pi) = \sum_{i=1}^{n=1} d_{\pi_i \pi(i+1)} + d_{\pi(n)\pi(1)} \qquad \pi \in \Pi\{1,2,\dots,n\} \tag{1}$$

The GA helps to replace the old ant's generation with the new one. Equation 2 describes the probability that ant k, located at node i, moves to node j, τ_{ij} is the pheromone level of edge (i, j), all taken at iteration t, and N_i is the set of one step neighbors of node i. While traversing an edge (i, j), the ant puts some pheromone on it, and the pheromone level of edge (i, j) is updated according to the following rule: $\tau_{ij}(t) \leftarrow \tau_{ij} + \Delta\tau$ where τ_{ij} is the iteration counter and $\Delta\tau$ is the constant amount of pheromone deposited by the ant.

$$P_{ij}^{k}(t) = \begin{cases} \dfrac{\tau_{ij}(t)}{\sum_{j \in N_i} \tau_{ij}(t)} \\ 0 \end{cases} \tag{2}$$

The best-expected solution should come in line with the short path length, driven by the process of releasing pheromones. Because the pheromone evaporates with time, links in longer routes will eventually contain much less pheromone linking the shorter tours.

Suppose n cities, with path p_{ij} between every pair of cities i, j. The length of p_{ij} is the distance between i and j, d_{ij}. The goal is to find path $P = p_{i_n i_1}$ where $(i_{1\dots n})$ is a permutation of $(1; :::; n)$ to find out the shortest path to visit the cities exactly once time.

III. Related Work

This section presents some relevant literature in the area. Harada et. al. [5] surveyed the works in the domain of parallel genetic algorithms.

Li et. al. [6] proposed a multiobjective evacuation route assignment model to plan an optimal egress route set for the individual evacuees to minimize the total evacuation time, minimize the total travel distance of all the evacuees and minimize the congestion during the evacuation process.

Wang et. al. [7] used a GPU-adapted Parallel Genetic Algorithm to solve the problem of generating daily activity plans for individual and household agents.

Yang et. al. [8] proposed an Electric Vehicle (EV) route model considering the fast charging and regular charging under the time-of-use price in the electricity market. The proposed model aims to minimize the total distribution costs of the EV route while satisfying the constraints of battery capacity, charging time and delivery/pickup demands, and the impact of vehicle loading on the unit electricity consumption per mile.

Yang et. al. [9] proposed cooperative scheduling rules and defined the overlapping time between the accelerating and braking trains for a peak-hours scenario and an off-peak-hours scenario, respectively. They also formulate a programming model to maximize the overlapping time with the headway time and dwell time control. They designed a GA with binary encoding to solve the optimal timetable. In [10], they formulated a two-objective integer programming model with headway time and dwell time control. Second, we design a genetic algorithm with binary encoding to find the optimal solution.

Tsai et. al. [11] presented an algorithm for reducing the computation time of GA and its variants using the traveling salesman problem.

Wang et. al. [12] introduced the untwist operator to improve the performance of GA to shorten the length of the route and quicken the convergent speed.

IV. Proposed Model

This section presents the proposed model, including its structure and parameters, and describes the algorithm driving it.

The proposed model is applied in two stages, exploring the space of solutions to produce two different outputs. The first stage identifies a set of activities to be assigned to workers, supported by the GA algorithm.

The second stage optimizes the route for the set of activities selected in the first stage, supported by the use of the ACO algorithm. This work is inspired by observing the foraging behavior of ant colonies [13].

Several experiments evaluate the model's effectiveness and explore alternative configurations to produce the best results in the selection stage. Three experiments in the first stage try to find out the best selection operator from the set including *roulette*, *remainder*, *uniform*, *tournament* and *stochunif.*

The final experiment evaluates the outcomes from the second stage and the performance of the overall model. For that purpose, the different solutions are evaluated according to the number of activities assigned to workers and the number of kilometers to execute them.

A. Experimental Setup

This section presents the experimental setup with its data structures, datasets, and parameters.

Several structures, including matrices and vectors, helped explore and validate the model's behavior with short datasets generalizing their application to larger datasets. The matrix "CoordinatesWorkOrders" maintains the coordinates for activities to be scheduled. "Workers" is the matrix with the reference for workers and coordinates of worker homes.

The "activities" is the vector with references for all the activities to be scheduled by the model.

The parameters drive the GA algorithm at the first stage, aiming to identify the optimized set of activities and the ACO optimizing the path to execute those activities at the second stage.

The first stage of the model gathers the required parameters by the GA to implement some of the business rules. In the case of `ACTIVITIES_TO_EVAL` constrains the maximum number of activities to be executed by workers while computing feasible solutions. The number of genes `NUMBER_GENES` corresponds to the number of activities to be encoded by in matrix "coordenatesWorkOrders". Four genes were used to support the encoding until a maximum number of 16 different activities (1111). One of the critical issues to cope with the model is the travel time by road between two locations. The average speed parameter with kilometers per hour (`AVERAGE_SPEED`) is used for that purpose. The `COEFFICIENT_TOUR` parameter normalizes the Euclidean distance between Global Positioning System (GPS) coordinates in order to have an approximation of the distance by road. Attending the fact that distance by road is not euclidean, a coefficient can help to reduce the error. The `PENALTY` is the core parameter driving the score of the fitness function of different other parameters aiming to implement other heuristics. The `MIN_WORK_TIME` is the parameter controlling the minimum work time for workers and impacts negatively the fitness score by `PENALTY/10`. The working hours per day should fit in the range from `MIN_WORK_TIME` to `MAX_WORK_TIME`. In order to have control over the total number of daily activities assigned to a given worker in a single workday, the fitness function penalizes the solutions according to `MAX_WORK_HOURS`, set equal to `PENALTY/100`. The parameter `ACTIVITY_TIME` sets the duration of all the different activities in the model. The `MIN_ACTIVITIES` parameter penalizes over the minimum number of activities or locations set with penalty `PENALTY/20`. A solution including repeated activities should be classified as not valid and penalized according to `PENALTY`. Parameter `MIGRATION_FRACTION` sets the percentage of individuals returning the best scores to migrate to the next generation. The crossover fraction, defined by the parameter `CROSSOVER_FRACTION`, corresponds to the fraction of genes swapped between individuals. Parameter `ELITE_COUNT` sets the percentage of the best individuals surviving for the next generation without any change. The elite count is computed according to the product `ELITE_COUNT` with the size of the population.

The second stage of the model tries to find out the best path over the activities by using the ACO. The `MAX_I_TIME` parameter sets the maximum execution time in minutes. The `NUMBER_ANTS` parameter sets the number of ants included in the simulation and corresponds to the number of activities.

The configuration of the ACO parameters sets generally adopted values: alpha (pheromone influence factor) equals 1 , beta (heuristic information importance) equals 5 , and rho (pheromone evaporation coefficient) equals 0.65 . The parameter PLOT enables the graphical presentation of the progress of the model, including the GA and ACO algorithms.

Table I gathers the parameters and their settings in experiments.

TABLE I. Parameters Settings

Configuration	Value
ACTIVITY_TIME	1.5
ACTIVITIES_TO_EVAL	15
AVERAGE_SPEED	60
COEFFICIENT_TOUR	1.3
CROSSOVER_FRACTION	0.1
ELITE_COUNT	0.05
GENERATIONS	100
MAX_I_TIME	1000
MAX_WORK_TIME	9
MIGRATION_FRACTION	0.1
MIN_ACTIVITIES	3
MIN_WORK_TIME	8
NUMBER_ANTS	1000
NUMBER_GENES	4
PENALTY	10000
POPULATION_SIZE	1000
PLOT	1

B. Algorithm

The following algorithm defines the steps to lookup for a optimized solution S^2 including the activities $a_i \in A$ assigned to workers $w_i \in W$, according to the following steps:

Algorithm 1: Model (P_{ga}, P_{aco}, $W, A, F, N, \Phi, \gamma, \alpha, \beta$)

INPUT:

P_{ga}, GA Parameters
P_{aco}, ACO Parameters
W, Workers
A, Activities
F, GA Fitness Function
N, Number of activities
Φ, Sort the activities by distance
γ, Extracts N activities
α, Extracts location
β, Computes Distance

for all $w_i \in W$ **do**
  for all $a_i \in A$ **do**
    $L_{w_i} \leftarrow \alpha(w_i)$
    $L_{a_i} \leftarrow \alpha(a_i)$
    $D_{w_i,a_i} \leftarrow \beta(L_{w_i}, L_{a_i})$
  end for
  $O^1_{w_i} \leftarrow \Phi(D)$
  $O^0_{w_i} \leftarrow \gamma(O^1_{w_i}, N)$
  for all $a_i \in O^0_{w_i}$ **do**
    $S^0_{w_i} \leftarrow S^0_{w_i} \cup a_i$
  end for
  $S^1_{w_i} \leftarrow GA(S^0_{w_i}, F, P_{ga})$
  $S^2_{w_i} \leftarrow ACO(S^1_{w_i}, P_{aco})$

OUTPUT: S^2, List of activities assigned to workers

C. GA Solution Encoding, Decoding and Fitness Function

This work adopted a binary scheme to encode individuals in the population into binary chromosomes (`POPULATION_SIZE`). Each individual denotes a candidate solution, carrying out a set of chromosomes as the set of activities assigned to workers. Each

individual represents a activities assigned to workers. Each individual represents a the worker.

An individual gathers 60 chromosomes to encode a set of 15 (`ACTIVITIES_TO_EVAL`) different activities while encoding each requires four chromosomes. Thus, a solution gathering the activities identified by indexes 2, 6, 4, 1, 3 and 8 can be encoded as an individual as follows: [2 6 4 1 3 0 0 0 0 0 0 0 0 0 8] is codified in binary as [[0010] [0110] [0100] [0001] [0011] [0000] [0000] [0000] [0000] [0000] [0000] [0000] [0000] [0000] [0111]]. Is also important to realize that alternative solutions such as [2 6 4 1 3 0 0 0 0 0 0 0 0 8], [3 4 0 0 0 8 0 6 0 1 0 0 2 0] and [2 4 0 0 0 8 0 6 0 1 0 0 0 3] have the same score. In such encoding, scheme value 0 has no meaning.

V. Experimental Work

This section presents the experimental work comprising a set of experiments evaluating the effectiveness of the proposed model and describes the structure of the solution and datasets. It discusses the implementation aspects, the involved datasets, and the settings.

A. Experimental Setup

MATLAB provided the package tools supporting the implementation of the proposed model with Genetic and Ant Colony algorithms. The quality of the solutions provided by individuals is scored by the GA fitness function of its Global Optimization Toolbox to lookup for a global minimum.

Table II presents the list of implemented MATLAB modules and describes their role.

TABLE II. MATLAB Model Modules

Module	Role
TELCO	Core application
FitnessTELCO	Fitness function
ACOTELCO	ACO implementation
Lldistkm	Distances in kilometres
DecodeWorkOrder	Decodes a binary solution
Assesstour	Validates a solution

The parameters settings adopted along the several experiments are defined according to the following Table I. Several experiments have been taken to explore these settings returning, including the maximum number of work hours" and "maximum activities to visit". The Fitness function $F()$ scores the best solutions as the ones not including repeated activities.

Fig. 1 depicts the progress with the scores of the fitness function along 100 generations and the information regarding the stall able to stop the algorithm's execution.

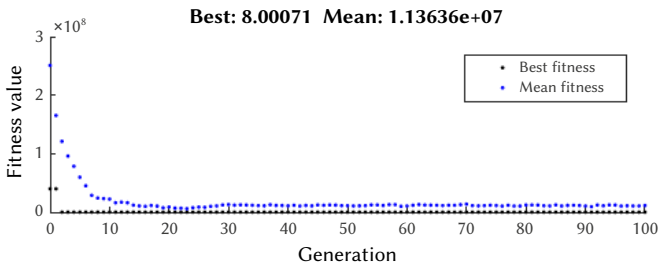


Fig. 1. Genetic Algorithm Evolution.

B. Datasets

Datasets for workers and activities are inputs to the model to compute the solutions. They include, respectively, the locations for worker base stations and locations where activities take place as GPS coordinates.

To support the following four experiments were used two different datasets. The purpose of the first dataset was to explore the effectiveness of the settings for the parameters containing 133 activities and 13 workers. To that aim, three experiments explored the most suitable GA operators in the selection of the activities along the first stage of the model.

A second larger dataset contains 436 activities and 191 workers. The fourth experiment uses this dataset to compare the quality of the solution produced by the proposed model with those produced by rule-based heuristic algorithms.

C. Experiments

The implementation of the proposed model is supported by the MATLAB toolbox environment providing EA. In order to demonstrate the capabilities for solving either complex or significant problems in the second stage of the model supported by the use of an ACO algorithm, a solution was computed from the second dataset according to Fig. 2.

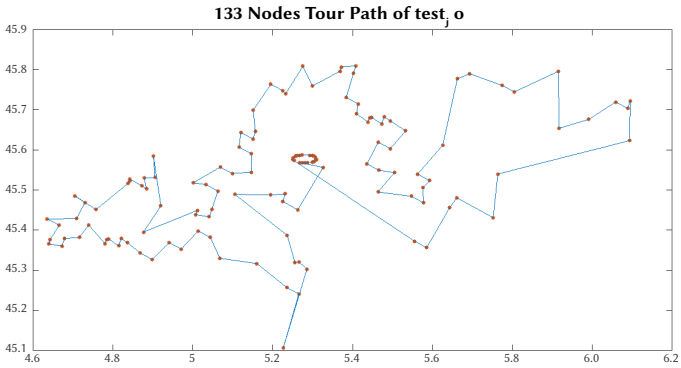


Fig. 2. ACO Scheduling Solution.

1. Experiment One

This experiment investigated the settings regarding the first stage of the proposed model. It relies on the use of the GA, including the genetic operator, the size of the population (`POPULATION_SIZE`), and the number of generations `GENERATIONS`. To that aim, the first dataset helped to explore the effectiveness of five different genetic selection operators.

The time to complete this experiment was 29338.619718 seconds (almost 8 hours). Table III summarizes the statistical results for the best GA scores for operators. From these results, it was possible to conclude that the tournament operator provided the best scores while reducing the number of individuals has a significant impact on the running time.

TABLE III. Experiment One Statistics

GA Operator	Mean	Min	Max	Std Dev.
Roulette	9.7229	8.0451	19.7617	3.4804
Remainder	10.6456	8.0411	23.8486	4.5726
Uniform	17.9256	8.0411	79.7528	16.2499
Tournament	9.4802	8.0414	79.7528	3.2803
Stochunif	9.8770	8.0576	19.9788	3.6922

Reducing the population (`POPULATION_SIZE`) from 1000 to 100 and the generations from 100 to 20 (parameter GENERATIONS significantly reduced the running time to 3825.188944 seconds (almost 1 hour).

2. Experiment Two

Departing from the already identified Tournament operator in the first experiment, this second one explored the effectiveness of its use with the GA. The number of generations gENERATIONS was set to 100 , and the first dataset was kept.

So, as a result, the score for the average value increased to 10.1522 while the computing time took 25838.210946 seconds (almost 7 hours). Table IV summarizes the achieved results.

TABLE IV. Experiment Two Statistics

GA Operator	Mean	Min	Max	Std Dev.
Tournament	10.1522	8.0406	28.2251	4.2976

3. Experiment Three

A third experiment explored the ability of the model to produce an effective plan in the shortest possible time. Thus, the first dataset (133 activities and 13 workers) and Tournament operator was selected, while the size of the population (`POPULATION_SIZE`) was reduced from 1000 to 100 individuals.

This experiment took 858.045974 seconds to run. Table V summarizes the statistical results from this experiment.

One of the key aspects regarding the quality of the solution comprises the number of activities assigned to workers. In this case, the number of assigned activities was 78 , and the non-assigned activities were 55 . Therefore, the ratio for activities assigned per worker was 6.

TABLE V. Experiment Three Statistics

GA Operator	Mean	Min	Max	Std Dev.
Tournament	8.0041	8.0000	8.6392	0.1753

4. Experiment Four

This last experiment explored a larger space of solutions and compared the results of this model with the ones from the heuristic rule-based model. For that purpose, a second dataset was used while the configurations from the experiment were maintained. As already stated, the second dataset includes 436 activities and 191 workers.

Computing the results took 3658.5085 seconds (almost 1 hour) to assign all the activities to 71 workers (100%), resulting in 6.1408 activities per worker. The kilometers needed for executing the activities assigned to the workers was 33859, and the average of kilometers per worker was 476.8858. The scoring average was 179.7771, which is worst compared to the optimal value of 8, achieved in the previous experiment.

In the case of the rule-based model, the assigned activities were 95 for 21 workers, while the non-assigned activities were 341 (21.79%). The activities per worker were 4.52, and the kilometers to execute the activities was 3315 (184.17 kilometers per worker).

VI. Results

Fig. 3 denotes the evolution of the fitness function $F()$ along several generations. The high scores at the beginning rapidly decrease and converge to a global minimum. Therefore, this global minimum corresponds to the available working hours, denoted by parameter `MIN_WORK_TIME`. Fig. 4 depicts the scores for the individuals contained in a given generation. Fig. 5 depicts the standard normal density function from the function operator tournament. The observed median value was 9.4802 while the standard deviation was 3.2803. Fig. 6 depicts the GA fitness scores for 13 different workers. Fig. 7 summarizes the online analysis for the best, worst and mean scores. Fig. 8 presents the computed path from the ACO algorithm with the set of activities assigned to a worker.

VII. Discussion

The experimental work evaluated the proposed model aiming to optimize the plan of activities for TELCO sector supported by GA and ACO along their first and second stages, respectively. The proposed model explored the space of solutions to maximize earnings, trying to discover a large possible number of activities to be assigned to workers and minimizing costs in terms of distance to perform the activities.

In that regard, the GA algorithm always tries to assign the maximum number of activities to workers (earnings), even if the number of kilometers and time are high (costs). The best solution (high score) from the first stage, supported by the use of GA, was achieved within twenty generations (parameter GENERATIONS). The computation time increases linearly with the number of generations. The fitness function penalizes the candidate solutions when they include non-singular activities. The best solutions are the ones that score the minimum number of hours in a day of work (8).

The results from the first three experiments denote the GA selection

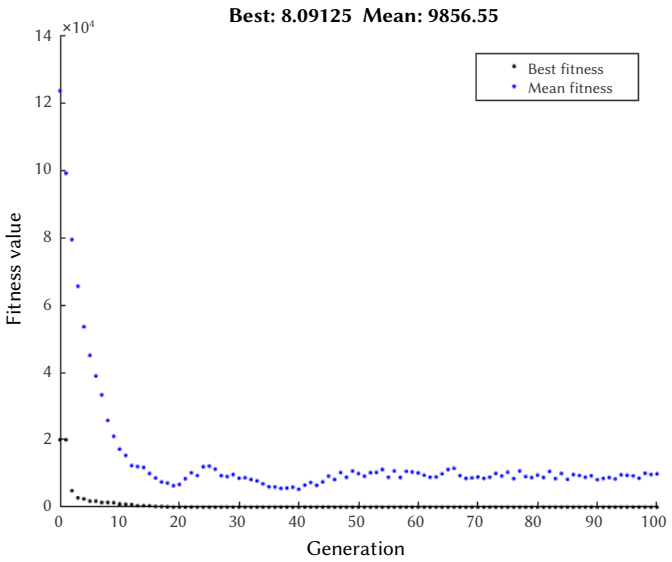


Fig. 3. Fitness Values Over Generations.

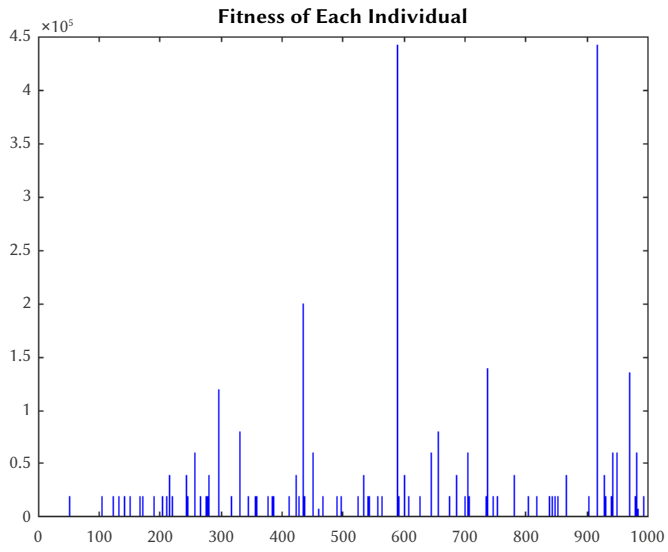


Fig. 4. Individual Fitness Function.

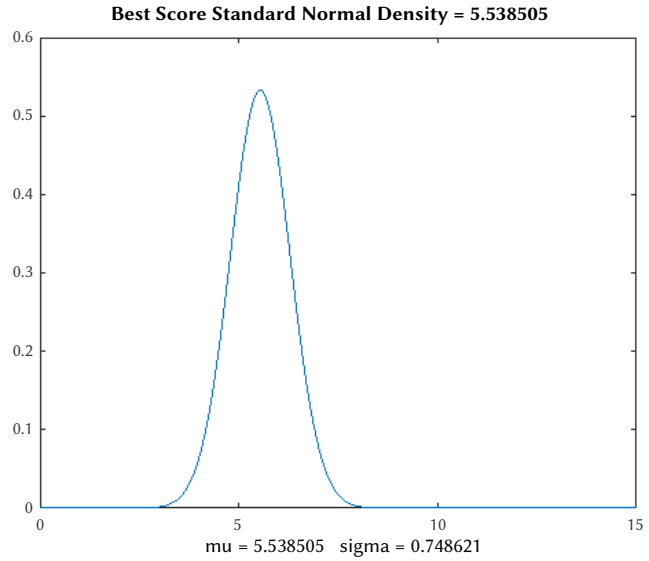


Fig. 5. Standard Normal Density Tournament Operator.

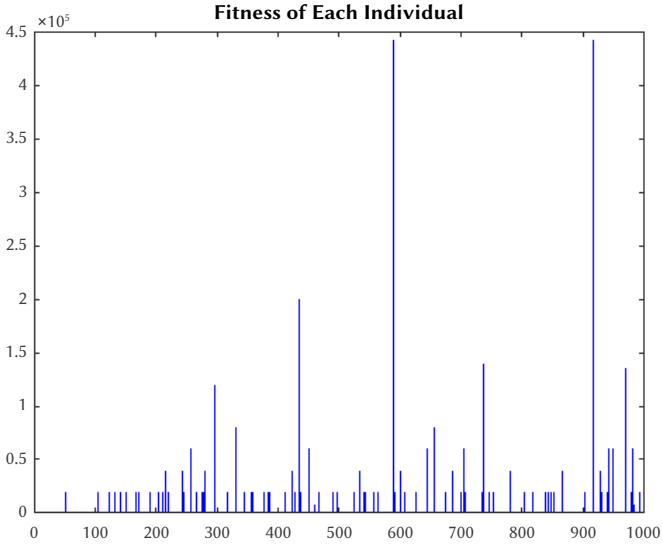


Fig. 6. GA Fitness Function Evolution.

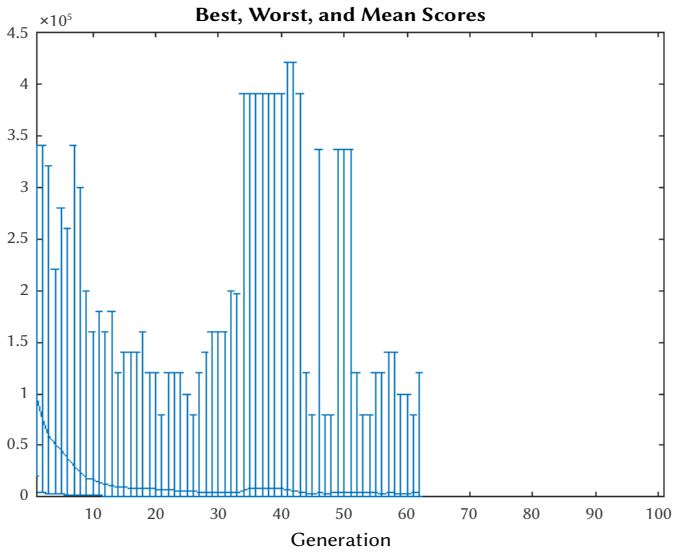


Fig. 7. GA Best Worst and Mean Values.

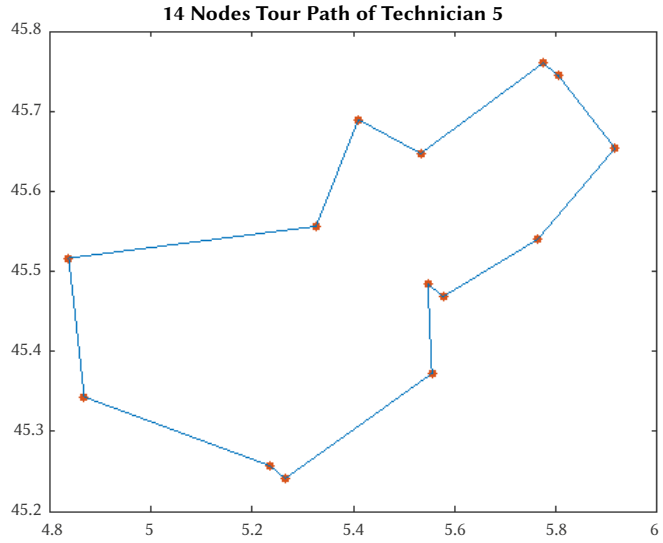


Fig. 8. ACO Activities Route.

operator tournament as the one providing the best scores. In addition, increasing the number of generations reduces the fitness function's average score. The best fitness score average was around 8 , according to the number of work hours in a single working day set by parameter `MIN_WORK_TIME`.

In the last experiment, a real dataset supports the comparison between the performance of the proposed model and the ruled-based model. All the activities were assigned to all the available workers, but at a cost. These results denote an acceptable solution despite the better performance of the ruled-based model regarding the ratio of activities per worker. It was also possible to depict a relevant number of workers assigned to a low number of activities. The cause is that the dataset for activities includes a significant number of activities distributed over vast regions.

From the analysis of experiment four, it was noticed that the distribution ratio per worker is significantly higher in comparison to the distribution provided by the rule-based model.

From the analysis of experiment four, it was noticed that the distribution ratio per worker is significantly higher in comparison to the distribution provided by the rule-based model.

One of the significant benefits of the model is that it does not require expert knowledge or a significant amount of effort in configuration activities to achieve an accurate solution. Thus, it is suitable to be applied at scale to larger datasets while requiring a minimum effort.

VIII. Conclusion

This work proposed a model inspired by nature to optimize the plan of the activities assigned to workers in the TELCO sector. A significant benefit of this approach comes from its effectiveness and reduced effort and time to compute a good solution, replacing the knowledge and work from experts. The model was supported by the use of Genetic Algorithms and Ant Colony Optimization Evolutionary Algorithms, and represents an alternative to the existing rule-based ones. The experimental results suggest that the model offers the foundation for its application in different use cases requiring the optimization of work-field services.

As the actual distance between the different locations is by road and not Euclidean, a regularization factor was introduced to diminuish the distance error, while future work can consider the use of distance. Future work will also pursue solutions for increasing the ratio of

assigned activities per worker. Moreover, future work will seek for a fitness function returning the earnings in Euros allowing an improved comparison in terms of quality to the different models.

Acknowledgment

This work was partially funded by the European Social Fund, through the Regional Operational Program Centro 2020, within the scope of the projects UIDB/05583/2020 and CISUC UID/CEC/00326/2020. Furthermore, we would like to thank the Research Center in Digital Services (CISeD) and the Polytechnic of Viseu for their support

References

[1] T. R. Cunha A. G., A. C. H., "Manual de computação evolutiva e metaheurística," in *Imprensa da Universidade de Coimbra, pp. 87-105.*

[2] E. Bonabeau, G. Theraulaz, J.-L. Deneubourg, S. Aron, S. Camazine, "Self-organization in social insects," *Trends in Ecology & Evolution*, vol. 12, no. 5, pp. 188–193, 1997.

[3] T. R. Cunha A. G., . Antunes C. H., "How trail laying and trail following can solve foraging problems for ant colonies," in *Behavioural Mechanisms of Food Selection, NATO-ASI Series, G20, Springer-Verlag, pp. 661–678.*

[4] M. Dorigo, T. Stutzle, "Ant colony optimization," in *MIT Press.*

[5] T. Harada, E. Alba, "Parallel genetic algorithms: a useful survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–39, 2020.

[6] Q. Li, Z. Fang, Q. Li, X. Zong, "Multiobjective evacuation route assignment model based on genetic algorithm," in *2010 18th International Conference on Geoinformatics*, 2010, pp. 1–5, IEEE.

[7] K. Wang, Z. Shen, "A gpu-based parallel genetic algorithm for generating daily activity plans," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1474–1480, 2012.

[8] H. Yang, S. Yang, Y. Xu, E. Cao, M. Lai, Z. Dong, "Electric vehicle route optimization considering time-of-use electricity price by learnable partheno- genetic algorithm," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 657–666, 2015, doi: 10.1109/TSG.2014.2382684.

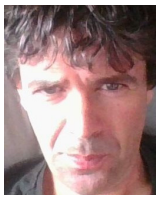
[9] X. Yang, X. Li, Z. Gao, H. Wang, T. Tang, "A cooperative scheduling model for timetable optimization in subway systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 438–447, 2012.

[10] X. Yang, B. Ning, X. Li, T. Tang, "A two-objective timetable optimization model in subway systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 1913–1921, 2014.

[11] C.-W. Tsai, S.-P. Tseng, M.-C. Chiang, C.-S. Yang, T.-P. Hong, "A high-performance genetic algorithm: using traveling salesman problem as a case," *The Scientific World Journal*, vol. 2014, 2014.

[12] L.-Y. Wang, J. Zhang, H. Li, "An improved genetic algorithm for tsp," in *2007 International Conference on Machine Learning and Cybernetics*, vol. 2, 2007, pp. 925–928, IEEE.

[13] Dorigo, C. D., "The ant colony optimization meta- heuristic," in *New Ideas in Optimization, McGraw-Hill, pp. 13–49.*



João Henriques

João Henriques is a PhD candidate in Science and Information Technology at the University of Coimbra (UC) and Adjunct Professor at the Department of Informatics Engineering at the Polytechnic of Viseu (IPV). His research interests at the Center for Informatics and Systems (CISUC) at C includes forensic and audit compliance for critical infrastructures protection. He also remains as Software Engineer in the private sector.

Filipe Caldeira

Filipe Caldeira is an Adjunct Professor at the Polytechnic Institute of Viseu, Portugal. He is a researcher at the CISeD research centre of the Polytechnic Institute of Viseu and at the Centre for Informatics and Systems of the University of Coimbra. His main research interests include ICT security, namely, trust and reputation systems, Smart Cities and Critical Infrastructure Protection. His research papers were published in various international conferences, journals and book chapters. He has been recently involved in some international and national research projects.

Edge Face Recognition System Based on One-Shot Augmented Learning

Diego M. Jiménez-Bravo[1,2*], Álvaro Lozano Murciego[1], André Sales Mendes[1], Luis Augusto Silva[1], Daniel H. De La Iglesia[1,3]

[1] Expert Systems and Applications Lab, Faculty of Science, University of Salamanca, Plaza de los Caídos s/n, 37008 Salamanca (Spain)
[2] Ontology Engineering Group, Departamento de Inteligencia Artificial, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Madrid (Spain)
[3] Faculty of Informatics, Pontifical University of Salamanca, 37002 Salamanca (Spain)

Received 6 April 2022 | Accepted 4 July 2022 | Early Access 12 September 2022

Abstract

There is growing concern among users of computer systems about how their data is handled. In this sense, IT (Information Technology) professionals are not unaware of this problem and are looking for solutions to meet the requirements and concerns of their users. During the last few years, various techniques and technologies have emerged that allow us to answer to the problem posed by users. Technologies such as edge computing and techniques such as one-shot learning and data augmentation enable progress in this regard. Thus, in this article, we propose the creation of a system that makes use of these techniques and technologies to solve the problem of face recognition and form a low-cost security system. The results obtained show that the combination of these techniques is effective in most of the face detection algorithms and allows an effective solution to the problem raised.

Keywords

Data Augmentation, Edge Artificial Intelligence, Edge Computing, Face Recognition, One-Shot Learning, Security System.

DOI: 10.9781/ijimai.2022.09.001

I. Introduction

Nowadays, we live in a society governed by technology and its advances, which generally aim to help society and make its day-to-day life easier and simpler. Large technology companies and in many cases public institutions are participants in these advances and put them into practice to obtain a business benefit in many cases. Society is aware of this benefit that companies and institutions can obtain and sometimes show rejection or mistrust of these innovations. Such systems can often reuse the information obtained from the user for other purposes, although this is not always the case. Given these circumstances, it is important to inform and educate users about these new technologies and how their information can be processed. Hence, this is an agreement in which companies and institutions have to participate with their commitment to being more transparent and, at the same time, users with the pro-activity to learn and understand the new systems.

With this in mind, one of the fields that most concerns society is the field of Artificial Intelligence (AI). In recent years, the advances developed in this field have been tremendous and its applications extend to practically all fields of society and in almost all areas of knowledge. Therefore, it is understandable that these systems, and especially the information they use and collect to provide results and benefits, can be a source of concern to users. In particular, when the data collected is related to them and can be used for other purposes or even sold to third parties. This collected data is commonly processed and stored in cloud systems. This is also where there is fear on behalf of users as they lose the perception of where their data is located.

Progress in computer science has led to the emergence of new computing techniques which are closer to the user, such as Fog and Edge computing. These techniques significantly reduce the amount of data sent over the Internet. However, the complexity of many AI systems and the number of resources they require make it impossible for such systems to run on devices with these computing architectures. Consequently, at present, it appears necessary to send data over the Internet to cloud services to obtain accurate and fast results in AI models with high computational needs.

However, currently, some techniques allow to reduce the amount of data sent through the Internet and they may reduce the amount of information obtained from the user or his environment (however, they still can collect huge amount of data from the users and theirs environments). However, these techniques, in which less data is used to train AI models, often have certain disadvantages associated with them, such as a decrease in the efficiency of the algorithms. Indeed, this makes perfect sense since complex models need large amounts of data to get good results, and if this data is reduced, the efficiency of the systems is reduced accordingly. However, it is important to note that despite this lack of data, many systems have been able to achieve high efficiency.

Unfortunately, this is not always the case; in fact, the vast majority

* Corresponding author.
E-mail address: dmjimenez@usal.es

of systems reduce their effectiveness. As a result, it is necessary to implement techniques that increase effectiveness despite the lack of data or that increase effectiveness even more despite a large amount of data. These techniques are known as data augmentation, which generates new data by applying slight modifications to the original data. In this way, the combination of data augmentation with the limited amount of user data can allow systems to achieve the minimum required performance and at the same time minimise the use of user data and/or information.

In this way, this paper presents the research developed by combining one-shot learning and data augmentation techniques to evaluate their effectiveness in systems that use fewer user data and also avoid sending data over the Internet. In this regard, the aim is to develop an edge computing system that makes use of an AI model developed through a combination of the two previous techniques for the intelligent recognition of different people.

A. Background

In this subsection, we analyze and explain the main concepts and techniques related to the areas of our research. Therefore, brief descriptions of each of the areas are included along with some of the most recent studies.

1. One-Shot Learning

The one-shot learning technique was first introduced in 2000 by Miller et al. [1] It is a subtype of supervised learning; as is well known, supervised learning uses labelled data or examples to learn and obtain knowledge and generate a model. Generally speaking, these types of training and models require a large amount of data to obtain good results. Nevertheless, the one-shot learning technique aims to do the opposite, with only one or a few samples it can train and generate an AI model [2]. However, this lack of data often has consequences in terms of model and results' efficiency.

An area where this technique has particular application is in the field of computer vision, where datasets with a large variety of samples are often not available. On the other hand, this area is one of the major applications of deep learning models; as is well known, these models require huge amounts of data to obtain highly efficient models. However, advances in science and research have led to the emergence of techniques such as transfer learning that allow the use of one-shot learning and obtain outstanding results.

The one-shot learning approach has been used in recent studies. Vinyals et al. [3] use it to increase the efficiency of other one-shot approaches on Omniglot and ImageNet datasets; specifically, they use matching nets to facilitate fast learning from a few labeled samples and classify images into the corresponding class. Another application is the use of one-shot learning for image segmentation [4]; they use a few labelled samples to extract the area of the image where the desired objects are located. This solution increases the efficiency of other similar investigations. Others, Woodward and Finn [5], have combined reinforcement learning and one-shot learning to increase the efficiency of pure supervised systems. They allow the systems to determine when is worth doing the classification or pay a penalty to receive the correct label. Another interesting approach is developed by Wang et al. [6] they combine the features extracted from the image along with the features located in the embedding of the class name. They mix these two pieces of information to map images and labels to predict unlabelled images. This approach obtains better performance than the baselines used in the research.

Moreover, one-shot learning techniques can also be applied to other areas. For example, [3] applies one-shot learning to text processing tasks, although not very successfully. In contrast, [7] applies this technique to drug discovery using a combination of LSTM (Long Short-Term Memory) and graph convolutional neural networks and a small number of samples.

2. Data Augmentation

As mentioned above, many models do not have the appropriate amount of data to train and implement a quality model. Furthermore, given the characteristics of these models, they are not able to obtain the insights needed from techniques such as one-shot learning. Thus, it is necessary to look for a solution capable of handling these two challenges. This solution is called data augmentation.

The main objective of data augmentation techniques is to increase the size and quality of the datasets. Like the previous technique, it is extremely related to the field of computer vision and imaging. Thus, image augmentation techniques [8] are the following: geometric transformations, color space augmentations, kernel filters, mixing images, random erasing, feature space augmentation, adversarial training, Generative Adversarial Networks (GAN), neural style transfer, and meta-learning. Generally speaking, these techniques can be classified into two main groups, (I) the most basic techniques, and (II) techniques based on deep learning solutions, especially those based on GAN architectures.

Besides, other more recent data augmentation techniques have emerged recently. A study developed by Zhong et al. [9] proposed to select a rectangle area of an image and erase its pixels with random values. This technique is called Random Erasing. On the other hand, Perez and Wang [10] proposed neural augmentation; this technique is based on a neural network that learns which data augmentation technique to use to maximise the performance of the system.

It can be observed that new data augmentation techniques are being proposed; in fact, new techniques are being proposed even outside the area of computer vision. This is the case of the study developed by Park et al. [11]; the researchers have developed SpecAugment, an augmentation method for automatic speech recognition.

3. Face Recognition

One of the most active areas of research in recent years in computer science is object recognition; one of its subfields is face recognition. This is a set of techniques that aim to identify faces within a set of images. The task of face recognition involves the development of other subtasks in a pipeline necessary for the correct performance of face recognition. These subtasks are listed and explained below:

1. **Face detection**: it determines the position and size of a human face in a digital image [12]. During the years several approaches have been presented, like Viola and Jones' Haar cascade method [13] to find face features with Haar-like features. A similar approach is Dlib HOG [14] which uses Histogram of Oriented Gradients (HOG) in the combination of Support Vector Machines (SVM) to detect faces. Dlib CNN [14] uses the power of Convolutional Neural Networks (CNN) to extract the features in faces; this is combined with Maximum-Margin Object Detector to maximize the results. Also, another well-known method is the SSD-Resnet [15], another CNN method to detect objects in digital images. More recent approaches like Multi-task Cascaded Convolutional Networks (MTCNN) [16] are also based on deep learning. MTCNN uses three subnetworks, P-Net, R-Net, and O-Net. These subnets are responsible for producing proposed regions, refining the proposals, and do face landmarking. Similarly, FaceNet [17] follows a process and uses CNN to detect a face on images.
2. **Face alignment**: it is the process in which the face landmark is usually rotated to obtain a face perfect alignment in case the face was originally rotated. Through the years several methods have been used to resolve this problem, ASMs [18], AAMs [19], CLMs [20], and cascade regression models [21]. Nonetheless, more recent

methods based on CNNs have been presented. These methods are divided into two main categories, coordinate regression models and heatmap regression models. A few examples of the first category are [22], [23]. On the other hand, examples of the second category are [24]–[26].

3. **Face encoder**: it is the process in which a face is transformed into an array representation. This array contains the features of the face and it is a numeric representation of it. This array of features is what usually is stored in a system that makes use of face recognition with its users. There are several techniques normally used to obtain the face features, such as LBPH (Local Binary Patterns Histogram) [27]; this technique divides the face into different blocks and obtains a histogram for each block, after that, it combines the histograms in only one. Another solution is OpenFace [28], a toolkit for face recognition. Also, ResNet [29] improves the encoding process as well as the whole process. Moreover, FaceNet [30] can also accomplish the encoding process.
4. **Face classifier**: this is the last stage of the process and consists of classifying the encoders from the previous phase into one of the classes provided during the classifier training or as an unknown person. Supervised learning algorithms are generally used. Thus, at this last stage, we obtain the results of the process, a label, and the confidence of the label.

This process is followed in recent studies using face recognition for different purposes. Such is the case of the [31]–[34] studies, among others.

4. Edge Computing

According to the European Edge Computing Consortium (EECC) edge computing is a computing paradigm in which certain services run near or on the very devices that request them [35]. Among other advantages, this type of computing increases privacy and decreases network latency. However, let us briefly describe some of the main features of edge computing according to the EECC [35].

- **Security**: security is an important aspect of all computing systems and it is important to ensure security between communications in an edge network. In terms of privacy, edge computing-based systems increase the privacy of data as all data is processed on these devices and the ownership of the data is kept between the owners of the edge devices.
- **Real-time**: the responses of services hosted on edge systems generally offer a much faster response and decrease the waiting time for users. This is mainly caused by lower data traffic between devices.
- **Acceleration**: resource-intensive processes that require large amounts of resources to respond to users, such as AI processes, can speed up their responses by finding computing centers closer to the user.
- **Management**: another important feature is the management that can be performed over edge computing networks. Firstly, its architecture allows it to be fail-stable and new devices can be added easily and with little or no change to the network topology. And another fundamental advantage is that by not depending on external services, it is possible to have a great deal of reliability in its operation.

The edge computing paradigm is widely used, proof of which are the studies carried out in recent years in areas such as autonomous vehicles [36], smart cities [37], smart homes [38], and even security systems such as the one proposed by Dirgantoro et al. [39] which proposes a security system based on recognition through Generative Adversarial Networks.

Taking all this into account, this article proposes the use of one-shot learning, data paradigm, face recognition, and edge computing paradigm to create a security system that is as transparent as possible for the end-user and based on edge devices.

The rest of the article is structured as follows: section II describes the proposed system, materials and methods of this study; section III shows the results obtained in the research; section IV explains the main ideas obtained from the results analysis; and finally, section V talks about the study conclusions and future lines of research.

II. Methods

In this section, we will describe the main features of the system proposed in this article as well as its software and hardware characteristics. The section is divided into three subsections.

A. Proposed System

The proposed system is based on the four concepts explained in the subsection A. To explain in a more detailed and precise way the proposed system in the following subsections, each of its modules is illustrated in Fig. 1 providing an overall view of the system.

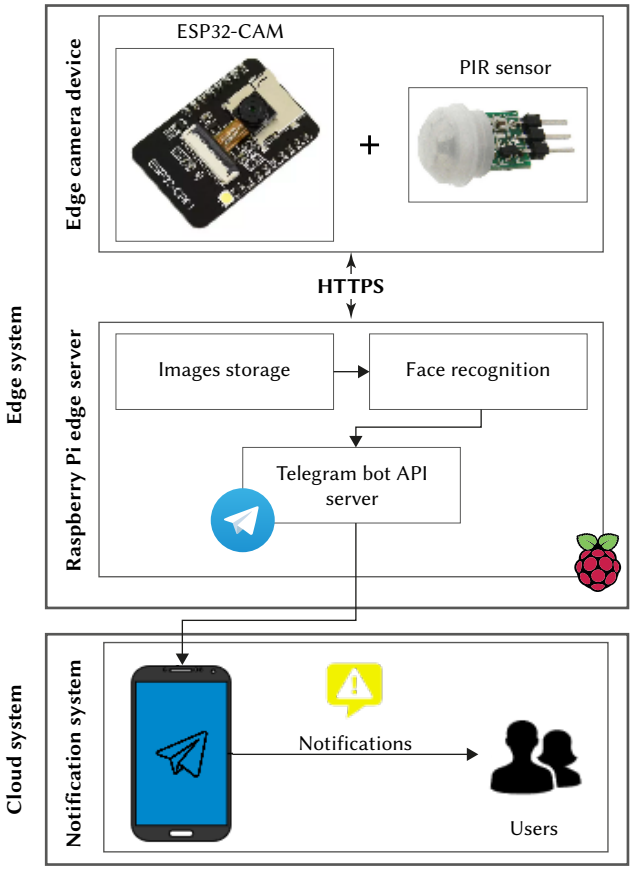


Fig. 1. Proposed system.

1. Edge Camera Device

As shown in Fig. 1, the system is composed of three main modules. The first one is the "Edge camera device", which is a hardware edge device that will take images whenever it detects movement in its proximity. Details of the specific hardware that composes this device will be explained in detail in 2. In addition, while the device detects movement, the system will take pictures every second to be able to detect in some of these images the faces of the people who have

caused the camera to wake up. In this way, this device will act as a mini edge security camera that will send the images obtained to another edge device.

2. Raspberry Pi Edge Server

The second module is the "Raspberry Pi edge server" which is formed by a Raspberry Pi (we will describe its features in the next section) that will act as a mini server to store both the images received by the edge camera device, the facial recognition model, and the Telegram bot in charge of notifying the user. The images are sent through the HTTPS protocol within a local network, so the images and the system aim to work as close as possible to the user and on devices that the user can have at home. On the other hand, the facial recognition model is trained from a single image of the users of the system using the one-shot augmented learning technique. In this way, the proposed system not only reduces the transit of user images over the Internet by using a local network but also reduces the number of user images required for the system to work. Furthermore, the models used to implement this model are proposed based on the study we implemented in this work where all possible combinations with the most relevant models for face recognition are studied (this part of the study is detailed in more detail in the next paragraph). Finally, this second module will also contain a space to store the API (Application Programming Interface) of the Telegram bot configured for our system. The bot will be in charge of using the facial recognition model to identify the faces in the images and program notifications to the Telegram account of the users when unknown people are identified in the images.

Face Recognition Study The facial recognition study was based on the techniques explained in A. With this study, we intend to test the effectiveness of the various existing models for face recognition with one-shot learning and one-shot augmented learning techniques. To carry out this study we propose different stages that allow us to obtain comparative results between the different methods. In Fig. 2 we can see a scheme of each of these stages and therefore of the general behaviour of this study. Likewise, this scheme is explained in detail below.

- **Data augmentation**: one of the objectives of the research is to compare the efficiency of one-shot learning and
- one-shot augmented learning techniques. To do so, we need to build two datasets for each of these techniques. The datasets are obtained from an original dataset (1) that is formed by folders associated with each of them to a person. To generate the one-shot augmented dataset, an image is selected and 100 new images are randomly generated from the selected one.
- The new images are generated by randomly modifying a series of variables: width and height shift range, brightness, zoom range, and rotation range.
- **Training datasets**: after the data augmentation process two datasets are obtained, the one-shot dataset obtained
- with the selected image and the one-shot augmented dataset obtained from the images generated from the image selected for the one-shot dataset.
- **Testing dataset**: the test set has the same structure as the previous datasets and contains the rest of the images not
- selected for the training sets. This set is used to test the resulting models after the training phase.
- **Face recognition**: the next step is to evaluate the techniques with each of the generated datasets. Therefore, a process is generated for the one-shot learning technique and another one for the one-shot augmented learning technique. For each, different combinations of models are tested; in particular, face detector models, face encoder models, and different classification algorithms are combined. In this way, it is possible to evaluate

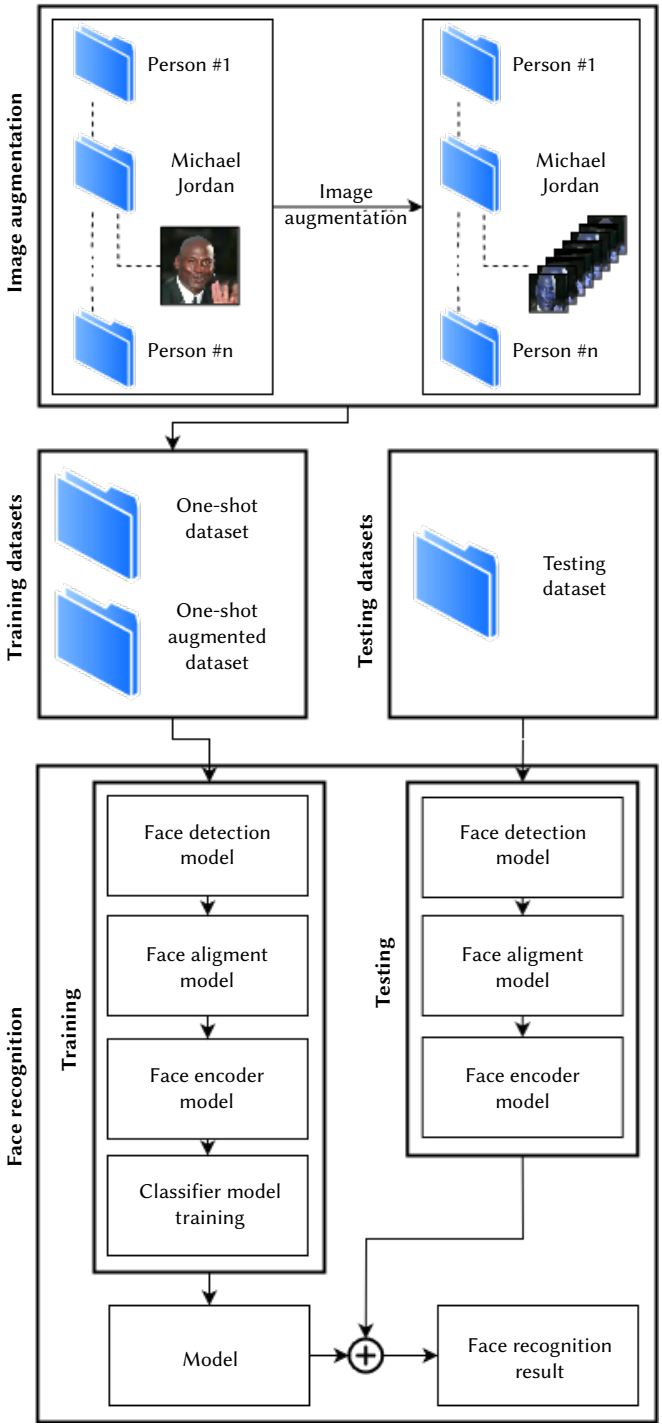


Fig. 2. Face recognition study proposal.

which combination of models works best with which dataset. The face detection models used are Haar Cascade (HC), Dlib HOG, Dlib CNN, SSD-Resnet, MTCNN, and FaceNet. On the other hand, the encoder models used are OpenFace, Dlib ResNet, and FaceNet. Finally, the classification models are as follows: Naive Bayes (NB), Linear Kernel SVM (LKSVM), Radial Basis Function (RBF) kernel SVM, k-Nearest Neighbours (k-NN), Decision Tree (DT), Random Forest (RF), Neural Network (NN), AdaBoost (AB), and QuaDrAtic (QDA) classifier.

After the training process, the resulting models are evaluated using the test set. The metrics used to evaluate these methods are explained in 2.

This process is repeated for different configurations of the datasets in which the number of classes and therefore of people vary. The purpose of this is to see how the effectiveness of both techniques also varies according to the number of classes.

3. Notification System

The last module, "Notification system", is responsible for receiving notifications received by the Telegram bot on a device compatible with the Telegram application. In this way, the user can consult them whenever he/she wishes, together with the image in which an unknown individual/s has/have been detected by the edge security system proposed in this article.

B. Materials

In this subsection, we will explain the dataset and hardware used during the course of this research.

1. Datasets

For the comparative study of one-shot learning and one-shot augmented learning techniques, a well-known dataset prepared for facial recognition models has been used. This dataset is called The labelled Faces in the Wild (LFW) dataset [40]. The dataset contains an average of 2.03 images per person, which is a total of 13,233 images for a total of 5,749 people. The internal structure of the dataset is divided into folders, one for each of the persons included in the dataset. Within these folders, we will find at least one image of the person to whom the folder refers. These images have a series of characteristics in common, all of them have a size of 250 x 250 pixels and are in JPG digital format.

According to [40] each of the images, belonging to the LFW dataset, are obtained employing the Haar cascade method proposed by [13] and increasing the detection area by a factor of 2.2 in each of its dimensions. This is done to obtain a larger viewing area than that provided by the Haar cascade. In this way, each image contains only one face per image allowing us to better evaluate one-shot learning and one-shot augmented learning techniques.

2. Hardware

The hardware used for the development of the prototype used in this article is detailed below.

- **ESP32-CAM**: this is one of the main components of the proposed system. The ESP32-CAM (embedded microcontroller, Espressif Systems, Shanghai, People's Republic of China) is an embedded microcontroller that operates independently. It has WiFi and Bluetooth connectivity. It also has a small integrated video camera and a MicroSD slot. Some other features of this microcontroller are as follows:
  - Connectivity: WiFi 802.11b/g/n, and Bluetooth 4.2 with BLE. Supports image upload over WiFi.
  - Connections: UART, SPI, I2C, and PWM. It has 9 GPIO pins.
  - Clock frequency: up to 160Mhz.
  - Microcontroller computational power: up to 600 DMIPS.
  - Memory: 520KB SRAM, 4MB PSRAM, and SD card slot.
  - Extras: has multiple sleep modes, firmware upgradeable by OTA, and LED for flash memory built-in.
  - Camera: supports OV2640 cameras that can be bundled or purchased separately. This type of camera has:
    - 2 MP sensor.
    - UXGA array size of 1622×1200 px.
    - Output format YUV422, YUV420, RGB565, RGB555 and 8-bit data compression.
    - It can transfer images between 15 and 60 FPS.
- **PIR sensor AM312**: it is a Passive Infrared Sensor (PIR sensor, ARCELI) that can detect the presence of moving objects in its area of action. Some of its characteristics are the following:
  - Working voltage: DC 2.7-12V.
  - Static power consumption: <0.1mA.
  - Delay time: 2 seconds.
  - Blocking time: 2 seconds.
  - Trigger: repeatable.
  - Detection range: cone angle of ď100 degrees, 3-5 m (required depending on the lens).
  - Working temperature: -20 to + 60.
  - Size PCB: 10 mm x 8 mm.
  - Overall size: Approx. 12 mm x 25 mm.
  - Lens Module: Small lens.
- **10k ohm electrical resistor**: 10k ohm electrical resistor (electrical resistor, AZ-Delivery Vertriebs GmbH, Deggendorf, Germany).
- **1k ohm resistor**: a 1k ohm electrical resistor (electrical resistor, AZ-Delivery Vertriebs GmbH, Deggendorf, Germany).
- **Transistor 2N3904**: a transistor (transistor, BOJACK, Guangdong, People's Republic of China) in charge of amplifying the signal coming from the PIR sensor AM312 and transmitting it to the ESP32-CAM.
- **Raspberry Pi 4 Model B**: the Raspberry Pi (computer board, Raspberry Pi, Cambridge, United Kingdom) will serve as a local server to host both the images transmitted by the ESP32-CAM and the Telegram bot that will be in charge of notifying the user. The technical characteristics of the model used are as follows:
  - RAM: 8GB.
  - Type of RAM: DDR3 SDRAM.
  - Operating system: Raspbian OS.
  - Processor: A-Series Dual-Core A4-3305.
  - Hardware interface: Bluetooth 5.0, WiFi 802.11b/g/n, Ethernet, Micro-HDMI, USB-C, USB 3.0, and USB 2.0.
  - Graphics card: Radeon Vega 8.
  - Graphics memory type: DDR4 SDRAM.
  - Graphics card interface: PCI-Express x4.
  - Voltage: 5 Volts.

C. Methods

In this subsection, we will explain all the methods and processes that have been followed during this research in such a way that the whole process is reproducible.

1. Materials Manipulation

Generation of Datasets and Subsets Generation of datasets and subsets: The dataset used for this research has been explained above (1). However, this dataset is modified to obtain two different datasets, one for the evaluation of the one-shot augmented dataset and the other for the evaluation of the one-shot dataset. Thus, the process to obtain these two datasets is explained below.

Most of the folders (people) of the LFW dataset contain only one image; in our experiment, these folders are not taken into account for the evaluation of the methods, since no different image would be available for the testing process. Therefore only folders containing more than one image will be taken into account in our experiments. From these folders we select a random image that will form the training subset for the one-shot dataset; on the other hand, the rest of the images in this folder will form the test subset for the one-shot

dataset. Once the one-shot learning dataset is formed, the process explained in 2 can be followed to generate the one-shot augmented learning dataset. To obtain the augmented images of the one-shot augmented learning dataset, the parameters and intervals described in Table I have been used.

TABLE I. Parameters Used for Data Augmentation

Parameter	Interval/value
Width shift range	[-1, 1]
Height shift range	[-1, 1]
Brightness range	[0.7, 1.3]
Zoom range	[-9, 11]
Rotation range	0.4

This methodology allows us to obtain two different datasets originating from the same input, the one-shot dataset and the one-shot augmented dataset. However, to evaluate the effectiveness of both techniques, several datasets have been configured for one-shot learning and one-shot augmented learning. The datasets are obtained in the same way but differ in the number of classes/people. As the number of classes decreases, the number of images available for each person is considered; in such a way that priority is always given to those that have more images that can be used for the test stage. As a result, pairs of datasets (one-shot and one-shot augmented) with 143, 85, 57, 41, and 19 classes are formed.

Edge Hardware Configuration The hardware components described in 2 constitute the different subsystems of the proposed system in A. In this way, we are going to explain how to configure each of these subsystems.

- **Edge camera device**: it consists of an ESP32-CAM, an AM312 PIR sensor, a 10k ohm electrical resistor, a 1k ohm electrical resistor, and a 2N3904 transistor. The electronic schematic of this subsystem is shown in Fig. 3. Also, keep in mind that the ESP32-CAM will remain asleep until the PIR sensor detects motion, and then the ESP32-CAM will wake up and take the necessary pictures that will be sent to the Raspberry Pi edge server.
- **Raspberry Pi edge server**: in this case, the system consists of a Raspberry Pi 4 Model B with Raspbian OS. In this device, an HTTP server has been configured to receive and store the images sent by the edge camera device. In addition, this device is responsible for processing these images with a facial recognition model, which we will specify in later sections, and will send the results through a Telegram bot to warn the user of possible intrusions.

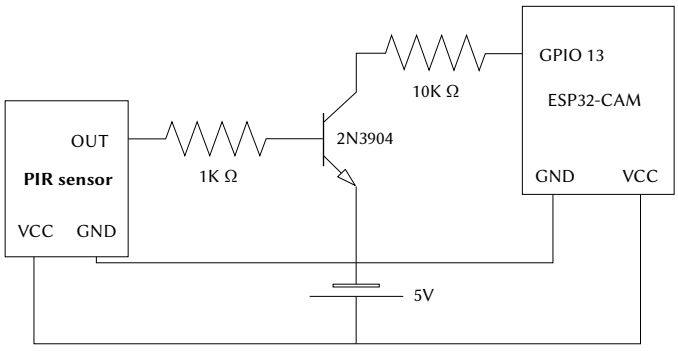


Fig. 3. Electronic scheme of the edge camera device.

2. Metrics

As explained above, the comparative study between one-shot learning and one-shot augmented learning techniques aims to obtain the best combination of models and therefore to establish the models to be used in the face recognition stage of the proposed system. However, the evaluation and comparison must be carried out taking into account some quality metrics. In this study, the accuracy metric described in (1) has been used for the evaluation:

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

where *TP* (True Positive) are the instances correctly classified as a positive class, *TN* (True Negative) are the instances correctly classified as negative instances, *FP* (False Positive) are the instances incorrectly classified as positive, and *FN* (False Negative) are the instances incorrectly classified as negative class.

III. Results

This section will explain the results of the different sub- systems or studies of the proposed system. This section will analyze and explain these results in an objective way.

A. Face Recognition Study Results

When analyzing the results of the face recognition study, different situations are observed, which we will show in this subsection. The totality of the results of each of the experiments can be seen in Table III which is included in the annex A.

Analyzing the results obtained and shown in Table III it can be observed that the one-shot augmented learning technique increases the efficiency of the combinations of the different algorithms concerning the one-shot learning. In the experiment with 143 classes, the one-shot augmented learning increases the results in 99 out of 162 algorithm combinations; in the experiment with 85 classes there are also 99 out of 162 algorithm combinations that improve their accuracy with the one-shot augmented learning technique; the experiment with 57 classes obtains better results in 95 algorithm combinations using the one-shot augmented learning; with 41 classes it improves in 94 algorithm combinations; while in the case study with 19 classes 110 combinations out of 162 increase their accuracy with the one-shot augmented learning technique.

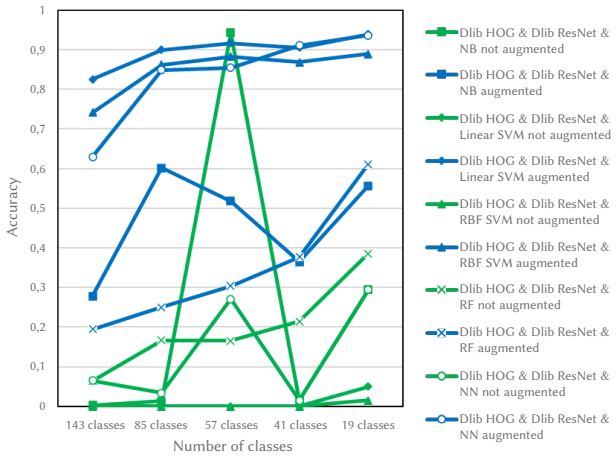


Fig. 4. One-shot augmented learning vs. one-shot learning.

However, it is important to mention that this increase in accuracy in many of the combinations is not significant. Nevertheless, there are certain combinations of algorithms especially those involving NB, Linear SVM, RBF SVM, RF, and NN classifiers. This can be seen in Fig. 4 where the evolution of these algorithms with the two techniques and in combination with the Dlib HOG and Dlib ResNet algorithms is illustrated. Thus, it can be seen that in general, these classifiers perform much better as they have more information to train their

(a) (b)

Fig. 5. Edge camera device location and view.

models. Similarly, and in general, for all the combinations studied, it can be observed that as the number of classes decreases, the accuracy of the algorithms increases.

Similarly, it is important to analyze which combination of algorithms is most effective in each of the experiments that have been developed. These can be found in Table II. It is interesting to note that the best combination of algorithms is the one formed by the Dlib HOG algorithm for face detection, the Dlib ResNet for the face encoder, and the k-NN algorithm as the face classification method. This combination of algorithms and methods obtains the best results in every experiment carried out in this study.

TABLE II. Best Algorithms Combinations

Experiment case	Algorithms combination	Accuracy value
143 classes not augmented	Dlib HOG & Dlib ResNet & k-NN	0.9096
143 classes augmented	Dlib HOG & Dlib ResNet & k-NN	0.8720
85 classes not augmented	Dlib HOG & Dlib ResNet & k-NN	0.9326
85 classes augmented	Dlib HOG & Dlib ResNet & k-NN	0.9201
57 classes not augmented	Dlib HOG & Dlib ResNet & k-NN and Dlib HOG & Dlib ResNet & NB	0.9427
57 classes augmented	Dlib HOG & Dlib ResNet & k-NN	0.9254
41 classes not augmented	Dlib HOG & Dlib ResNet & k-NN	0.9429
41 classes augmented	Dlib HOG & Dlib ResNet & k-NN	0.9102
19 classes not augmented	Dlib HOG & Dlib ResNet & k-NN	0.9477
19 classes augmented	Dlib HOG & Dlib ResNet & k-NN	0.9400

B. Hardware Systems Results

Following the design model of the hardware components shown in Fig. 3, an edge camera device that sends photos to the Raspberry Pi edge server every time it detects movement in the viewing area of its PIR sensor has been formed. This hardware is complemented with the software loaded in the ESP32-CAM microcontroller that is in charge of reading the signal received from the PIR sensor and waking up the ESP32-CAM module to capture and send the images to the edge server.

To protect the hardware, it is housed in a 3D printed casing so that its components are protected and less accessible to users. This housing will be located at the entrance door of a house as shown in Fig. 5a. In addition, a view of what the ESP32-CAM's optical sensor observes can be seen at Fig. 5b.

C. Notification Systems Results

The notification system consists of a Telegram bot that the user has to start in his Telegram application and from that moment on he will receive notifications when an unknown person approaches the edge device. This notification will provide a picture of the unknown person along with the exact time the image was taken. An example of the bot chat can be seen at Fig. 6.

D. System Recognition Results

The resulting system has been tested in a real environment to prove its effectiveness with images captured by the hardware developed during the research. As explained in B the hardware has been placed at the front door of a house. This way when a person walks into the hardware the PIR sensor wakes up the ESP32-CAM module and takes pictures which are sent to the Raspberry Pi edge server for analysis.

In this sense, we have performed tests to capture images that are analyzed by the Telegram bot and generated models based on the Dlib HOG detector, the Dlib ResNet encoder, and the k-NN classifier. Thus, the system has been tested with different lighting conditions and by placing the face in different positions.

During these experiments, we have trained a model based on a photograph of any of the people that the model will be able to identify (in our case a single person) that has been augmented following the data augmentation principles already followed previously. This model will be in charge of analyzing the images captured by the hardware.

After analyzing the results, it is observed that in the real environment the model obtains an accuracy value of 100%.

IV. Discussion

After observing the results shown in III, certain conclusions can be drawn that are interesting to discuss in this section.

In general, it can be observed that the one-shot augmented learning approach increases the results concerning the one-shot learning approach. However, in certain combinations of algorithms, the

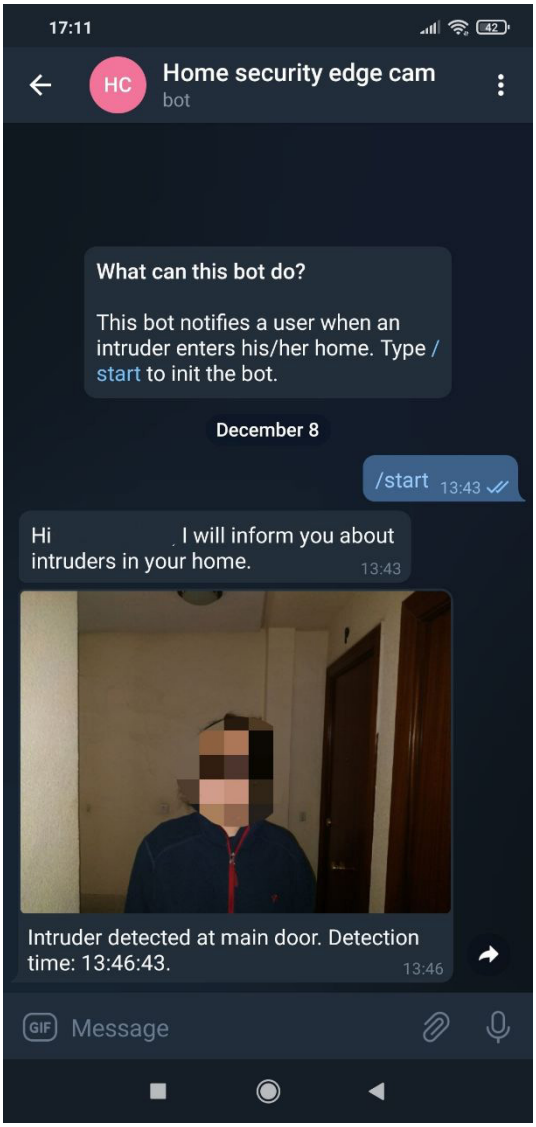


Fig. 6. Telegram bot chat screenshot.

difference is practically negligible, or in others, the results are lower. But in the same way, certain combinations of algorithms such as those shown in Fig. 4 greatly increase the results concerning one-shot learning. This makes us think that specifically these combinations of algorithms, especially the classifier, work better the more data variety they have.

Similarly, it can be seen that the best performing combination of algorithms is the combination of the Dlib HOG, Dlib ResNet, and k-NN algorithms. It is this combination of algorithms that has been used to identify the faces captured by the edge camera device. And as mentioned in D, the accuracy of the model is 100%. In principle, the expected efficiency of the model was high, although such a high efficiency was not expected. The efficiency is likely to be increased as the model has fewer classes and can discern better between users and because the position of the sensor allows good quality images to be obtained. The fact that the accuracy of the algorithm increases as the number of classes decreases is also supported by the results shown in Table 3; it can be observed that the efficiency and the number of classes are inversely proportional to each other.

Furthermore, in this study, it has to be taken into account that the use of edge and one-shot learning technology reduces one of the concerns of computer system users, the use of their data. The proposed system makes use of easily and cheaply available hardware that is always within the user's reach and connected to a local Internet network. On the one hand, the edge solution allows the user's data to be processed on devices close to the user and reduces the transit of information through cloud systems, with the user being the owner of his or her information. On the other hand, the use of one-shot learning reduces the amount of user information needed to train AI models.

On the other hand, the alert system tries to be as unobtrusive as possible for the user. First of all, using an instant messaging application such as Telegram allows the user to easily use the notification system. Secondly, the configuration of the Telegram bot itself prevents it from sending constant notifications to the user; only when a user unknown to the system approaches the edge camera device. Thus, it can be seen that the proposed system offers a comprehensive system for home security control using little data and powerful and accurate AI algorithms for the correct functioning of the alert system.

V. Conclusions

In this paper, we have built a face recognition system based on edge computing technologies and the combination of one-shot learning and data augmentation. The recognition system also has a notification system based on a bot of the Telegram application capable of notifying the user of intruders or unknown persons in the vicinity of their home.

In addition, the facial recognition system is efficient and capable of running on systems with limited computing capabilities. On the other hand, the effectiveness of the one-shot augmented learning technique has been demonstrated in certain algorithms focused on facial recognition. Moreover, we have discovered the perfect combination of detection, encoder, and classification algorithms for tasks in this domain.

Also, the fact that the system is focused on the edge computing paradigm allows users to be more aware of the information that is used about them in a local network environment. In this way, users are less reticent to use these technologies, and the security and speed in the delivery of results are improved concerning computing systems such as the cloud.

Thus, this article also demonstrates that complex artificial intelligence models can now be run on edge devices with certain computational capabilities without time and response delays. In the case of the proposed system, this is an important feature because the speed of intrusion detection and notification is essential for the physical security of the system's users.

On the other hand, following the conclusion of this study, new lines of research have emerged that may be interesting to address in the future. For example, it would be interesting to implement a face detection model for microcontrollers in such a way as to reduce the transit of images between the ESP32-CAM microcontroller and the Raspberry Pi edge server; in this way, the risk of the photos taken by the edge camera device being leaked to third parties can be minimized as much as possible, as only the identity of the person/s detected would be transmitted. Another interesting area of research is to analyze whether the one-shot augmented learning technique is equally effective in other areas of object detection, such as traffic analysis in urban environments.

Appendix

A. Face Recognition Study Results

We show in Table III a complete table with the results of all the experiments carried out in the facial recognition study.

TABLE III. Face Recognition Study Results

Models	143 classes		85 classes		57 classes		41 classes		19 classes	
	Not augmented	Augmented	Not augmented	Augmented	Not augmented	Augmented	Not augmented	Augmented	Not augmented	Augmented
Haar Cascade & OpenFace & NB	0.0034	0.0680	0.0105	0.1034	0.0086	0.1286	0.0138	0.1083	0.0329	0.2123
Haar Cascade & OpenFace & Linear SVM	0.0014	0.0737	0.0036	0.1125	0.0010	0.1385	0.0022	0.1337	0.0144	0.2327
Haar Cascade & OpenFace & RBF SVM	0.0030	0.0701	0.0094	0.0926	0.0090	0.1263	0.0080	0.1297	0.0284	0.2133
Haar Cascade & Open Face & k-NN	0.0996	0.0744	0.1368	0.1037	0.1651	0.1318	0.1472	0.1508	0.2197	0.2043
Haar Cascade & OpenFace & DT	0.0050	0.0107	0.0138	0.1205	0.0138	0.0205	0.0262	0.0523	0.0543	0.1091
Haar Cascade & OpenFace & RF	0.0261	0.0430	0.0312	0.0675	0.0473	0.0720	0.0494	0.0912	0.1405	0.1699
Haar Cascade & OpenFace & NN	0.0522	0.0614	0.0626	0.1114	0.0525	0.1219	0.0556	0.1312	0.0384	0.2063
Haar Cascade & OpenFace & AB	0.0055	0.0076	0.0176	0.1186	0.0230	0.0189	0.0229	0.0222	0.0653	0.0448
Haar Cascade & OpenFace & QDA	0.0055	0.0034	0.0176	0.0105	0.0230	0.0106	0.0229	0.0163	0.0653	0.0309
Haar Cascade & Dlib ResNet & NB	0.0062	0.3899	0.0105	0.6261	0.0080	0.6673	0.0313	0.6185	0.0404	0.7613
Haar Cascade & Dlib ResNet & Linear SVM	0.0002	0.7429	0.0011	0.8258	0.0000	0.8439	0.0004	0.8201	0.0015	0.8989
Haar Cascade & Dlib ResNet & RBF SVM	0.0005	0.7008	0.0011	0.7827	0.0006	0.8228	0.0004	0.7820	0.0055	0.8825
Haar Cascade & Dlib ResNet & k-NN	0.8679	0.8162	0.8889	0.8729	0.8861	0.8759	0.8870	0.8674	0.9008	0.9038
Haar Cascade & Dlib ResNet & DT	0.0098	0.0108	0.0179	0.0447	0.0534	0.0390	0.0560	0.0930	0.0597	0.1280
Haar Cascade & Dlib ResNet & RF	0.0994	0.1401	0.1238	0.2294	0.1398	0.1663	0.2013	0.4001	0.3478	0.5850
Haar Cascade & Dlib ResNet & NN	0.0469	0.4064	0.0163	0.7717	0.0077	0.7780	0.0403	0.8216	0.0344	0.9033
Haar Cascade & Dlib ResNet & AB	0.0085	0.0151	0.0565	0.0342	0.0528	0.0198	0.0465	0.0443	0.1151	0.0772
Haar Cascade & Dlib ResNet & QDA	0.0085	0.0041	0.0565	0.0119	0.0528	0.0150	0.0465	0.0185	0.1151	0.0274
Haar Cascade & FaceNet & NB	0.0034	0.2775	0.0105	0.2934	0.0102	0.3474	0.0149	0.3521	0.0229	0.4668
Haar Cascade & FaceNet & Linear SVM	0.0002	0.4519	0.0008	0.4921	0.0010	0.5691	0.0029	0.3899	0.0065	0.6517
Haar Cascade & FaceNet & RBF SVM	0.0046	0.2580	0.0025	0.3300	0.0016	0.2889	0.0022	0.2544	0.0060	0.5047
Haar Cascade & FaceNet & k-NN	0.3988	0.5501	0.5230	0.5966	0.5390	0.6372	0.5214	0.5385	0.6572	0.6841
Haar Cascade & FaceNet & DT	0.0112	0.0080	0.0099	0.0179	0.0198	0.0266	0.0291	0.0400	0.0862	0.0693
Haar Cascade & FaceNet & RF	0.0327	0.0888	0.0585	0.1977	0.0972	0.2505	0.0763	0.1867	0.2013	0.3672
Haar Cascade & FaceNet & NN	0.2772	0.4464	0.3303	0.5751	0.2879	0.5861	0.1439	0.4931	0.1061	0.7145
Haar Cascade & FaceNet & AB	0.0060	0.0050	0.0265	0.0168	0.0377	0.0157	0.0578	0.0164	0.1216	0.0304
Haar Cascade & FaceNet & QDA	0.0060	0.0451	0.0265	0.0540	0.0377	0.0499	0.0578	0.0730	0.1216	0.1420
Dlib HOG & OpenFace & NB	0.0083	0.0033	0.0233	0.0108	0.2370	0.0099	0.0409	0.0169	0.1096	0.0353
Dlib HOG & OpenFace & Linear SVM	0.0082	0.0033	0.0233	0.0108	0.0003	0.0099	0.0409	0.0169	0.1096	0.0353
Dlib HOG & OpenFace & RBF SVM	0.0083	0.0033	0.0233	0.0108	0.0003	0.0099	0.0409	0.0169	0.1096	0.0353
Dlib HOG & Open Face & k-NN	0.0083	0.0033	0.0233	0.0108	0.2370	0.0099	0.0409	0.0169	0.1096	0.0353
Dlib HOG & OpenFace & DT	0.0083	0.0033	0.0233	0.0108	0.0186	0.0099	0.0409	0.0169	0.1096	0.0353
Dlib HOG & OpenFace & RF	0.0083	0.0033	0.0233	0.0108	0.0620	0.0099	0.0409	0.0169	0.1096	0.0353
Dlib HOG & OpenFace & NN	0.0083	0.0033	0.0233	0.0108	0.1574	0.0099	0.0409	0.0169	0.1096	0.0353
Dlib HOG & OpenFace & AB	0.0083	0.0033	0.0233	0.0108	0.0334	0.0099	0.0409	0.0169	0.1096	0.0353
Dlib HOG & OpenFace & QDA	0.0083	0.0033	0.0233	0.0108	0.0334	0.0099	0.0409	0.0169	0.1096	0.0353
Dlib HOG & Dlib ResNet & NB	0.0035	0.2779	0.0134	0.6016	0.9427	0.5174	0.0154	0.3651	0.2946	0.5564

Models	143 classes		85 classes		57 classes		41 classes		19 classes	
	Not augmented	Augmented	Not augmented	Augmented	Not augmented	Augmented	Not augmented	Augmented	Not augmented	Augmented
Dlib HOG & Dlib ResNet & Linear SVM	0.0007	0.8246	0.0000	0.8993	0.0000	0.9165	0.0000	0.9046	0.0497	0.9390
Dlib HOG & Dlib ResNet & RBF SVM	0.0005	0.7410	0.0000	0.8615	0.0000	0.8820	0.0000	0.8681	0.0154	0.8888
Dlib HOG & Dlib ResNet & k-NN	0.9096	0.8720	0.9326	0.9201	0.9427	0.9254	0.9429	0.9098	0.9477	0.9400
Dlib HOG & Dlib ResNet & DT	0.0106	0.0097	0.0162	0.0466	0.0242	0.0484	0.0451	0.0793	0.1522	0.3509
Dlib HOG & Dlib ResNet & RF	0.0646	0.1943	0.1667	0.2502	0.1654	0.3033	0.2145	0.3768	0.3852	0.6112
Dlib HOG & Dlib ResNet & NN	0.0654	0.6298	0.0344	0.8487	0.2711	0.8542	0.0154	0.9102	0.2940	0.9355
Dlib HOG & Dlib ResNet & AB	0.0170	0.0168	0.0421	0.0358	0.1385	0.0282	0.0796	0.0237	0.1619	0.3279
Dlib HOG & Dlib ResNet & QDA	0.0170	0.0035	0.0421	0.0114	0.1385	0.0259	0.0796	0.0154	0.1619	0.2941
Dlib HOG & FaceNet & NB	0.0035	0.0378	0.0199	0.0478	0.4332	0.0447	0.0616	0.0616	0.1086	0.1296
Dlib HOG & FaceNet & Linear SVM	0.0035	0.0378	0.0199	0.0478	0.0007	0.0447	0.0616	0.0616	0.1086	0.1296
Dlib HOG & FaceNet & RBF SVM	0.0035	0.0378	0.0199	0.0478	0.0007	0.0447	0.0058	0.0616	0.1086	0.1296
Dlib HOG & FaceNet & k-NN	0.0035	0.0378	0.0199	0.0478	0.4332	0.0447	0.0616	0.0616	0.1086	0.1296
Dlib HOG & FaceNet & DT	0.0035	0.0378	0.0199	0.0478	0.0166	0.0447	0.0616	0.0616	0.1086	0.1296
Dlib HOG & FaceNet & RF	0.0035	0.0378	0.0199	0.0478	0.0620	0.0447	0.0616	0.0616	0.1086	0.1296
Dlib HOG & FaceNet & NN	0.0035	0.0378	0.0199	0.0478	0.3099	0.0447	0.0616	0.0616	0.1086	0.1296
Dlib HOG & FaceNet & AB	0.0035	0.0378	0.0199	0.0478	0.0308	0.0447	0.0616	0.0616	0.1086	0.1296
Dlib HOG & FaceNet & QDA	0.0035	0.0378	0.0199	0.0478	0.0308	0.0447	0.0616	0.0616	0.1086	0.1296
Dlib CNN & OpenFace & NB	0.0064	0.0032	0.0168	0.0108	0.0235	0.0116	0.0368	0.0149	0.0882	0.0304
Dlib CNN & OpenFace & Linear SVM	0.0064	0.0032	0.0168	0.0108	0.0234	0.0116	0.0368	0.0149	0.0882	0.0304
Dlib CNN & OpenFace & RBF SVM	0.0064	0.0032	0.0168	0.0108	0.0235	0.0116	0.0368	0.0149	0.0882	0.0304
Dlib CNN & Open Face & k-NN	0.0064	0.0032	0.0168	0.0108	0.0235	0.0116	0.0368	0.0149	0.0882	0.0304
Dlib CNN & OpenFace & DT	0.0064	0.0032	0.0168	0.0108	0.0234	0.0116	0.0368	0.0149	0.0882	0.0304
Dlib CNN & OpenFace & RF	0.0064	0.0032	0.0168	0.0108	0.0235	0.0116	0.0368	0.0149	0.0882	0.0304
Dlib CNN & OpenFace & NN	0.0064	0.0032	0.0168	0.0108	0.0235	0.0116	0.0368	0.0149	0.0882	0.0304
Dlib CNN & OpenFace & AB	0.0064	0.0032	0.0168	0.0108	0.0235	0.0116	0.0368	0.0149	0.0882	0.0304
Dlib CNN & OpenFace & QDA	0.0064	0.0032	0.0168	0.0108	0.0235	0.0116	0.0368	0.0149	0.0882	0.0304
Dlib CNN & Dlib ResNet & NB	0.0071	0.4392	0.0119	0.7280	0.0090	0.0116	0.0186	0.4529	0.3034	0.5989
Dlib CNN & Dlib ResNet & Linear SVM	0.0002	0.8041	0.0003	0.8695	0.0045	0.0116	0.0004	0.8892	0.0010	0.9108
Dlib CNN & Dlib ResNet & RBF SVM	0.0011	0.7619	0.0003	0.8381	0.0045	0.8525	0.0004	0.8240	0.0010	0.8804
Dlib CNN & Dlib ResNet & k-NN	0.8848	0.8429	0.9092	0.8891	0.9148	0.8994	0.9122	0.8899	0.9183	0.9118
Dlib CNN & Dlib ResNet & DT	0.0135	0.0112	0.0317	0.0345	0.0312	0.0437	0.0652	0.0612	0.1380	0.3473
Dlib CNN & Dlib ResNet & RF	0.0516	0.1489	0.1799	0.2720	0.1630	0.2713	0.2077	0.3407	0.3827	0.6846
Dlib CNN & Dlib ResNet & NN	0.1172	0.5890	0.0248	0.8378	0.0283	0.8271	0.0292	0.8750	0.3338	0.8954
Dlib CNN & Dlib ResNet & AB	0.0289	0.0048	0.0394	0.0270	0.0543	0.1356	0.0893	0.0361	0.3627	0.2890
Dlib CNN & Dlib ResNet & QDA	0.0289	0.0037	0.0394	0.0113	0.0543	0.0084	0.0893	0.0149	0.3627	0.2985
Dlib CNN & FaceNet & NB	0.0046	0.0328	0.0171	0.0436	0.0289	0.0437	0.0517	0.0714	0.0977	0.1305
Dlib CNN & FaceNet & Linear SVM	0.0046	0.0328	0.0171	0.0436	0.0289	0.0437	0.0517	0.0714	0.0977	0.1305
Dlib CNN & FaceNet & RBF SVM	0.0046	0.0328	0.0171	0.0436	0.0289	0.0437	0.0517	0.0714	0.0977	0.1305
Dlib CNN & FaceNet & k-NN	0.0046	0.0328	0.0171	0.0436	0.0289	0.0437	0.0517	0.0714	0.0977	0.1305

Models	143 classes		85 classes		57 classes		41 classes		19 classes	
	Not augmented	Augmented	Not augmented	Augmented	Not augmented	Augmented	Not augmented	Augmented	Not augmented	Augmented
Dlib CNN & FaceNet & DT	0.0046	0.0328	0.0171	0.0436	0.0289	0.0437	0.0517	0.0714	0.0977	0.1305
Dlib CNN & FaceNet & RF	0.0046	0.0328	0.0171	0.0436	0.0289	0.0437	0.0517	0.0714	0.0977	0.1305
Dlib CNN & FaceNet & NN	0.0046	0.0328	0.0171	0.0436	0.0289	0.0437	0.0517	0.0714	0.0977	0.1305
Dlib CNN & FaceNet & AB	0.0046	0.0328	0.0171	0.0436	0.0289	0.0437	0.0517	0.0714	0.0977	0.1305
Dlib CNN & FaceNet & QDA	0.0046	0.0328	0.0171	0.0436	0.0289	0.0437	0.0517	0.0714	0.0977	0.1305
SSD-Resnet & OpenFace & NB	0.0032	0.0659	0.0185	0.0107	0.0334	0.0105	0.0228	0.0186	0.0815	0.0349
SSD-Resnet & OpenFace & Linear SVM	0.0038	0.0032	0.0185	0.0107	0.0335	0.0096	0.0022	0.0186	0.0815	0.0349
SSD-Resnet & OpenFace & RBF SVM	0.0038	0.0032	0.0185	0.0107	0.0335	0.0105	0.0112	0.0186	0.0815	0.0349
SSD-Resnet & Open Face & k-NN	0.0038	0.0032	0.0185	0.0107	0.0335	0.0105	0.1317	0.0186	0.0815	0.0349
SSD-Resnet & OpenFace & DT	0.0038	0.0032	0.0185	0.0107	0.0335	0.0105	0.0231	0.0186	0.0815	0.0349
SSD-Resnet & OpenFace & RF	0.0038	0.0032	0.0185	0.0107	0.0335	0.0105	0.0465	0.0186	0.0815	0.0349
SSD-Resnet & OpenFace & NN	0.0038	0.0032	0.0185	0.0107	0.0335	0.0105	0.0519	0.0186	0.0815	0.0349
SSD-Resnet & OpenFace & AB	0.0038	0.0032	0.0185	0.0107	0.0335	0.0105	0.0167	0.0186	0.0815	0.0349
SSD-Resnet & OpenFace & QDA	0.0038	0.0032	0.0185	0.0107	0.0335	0.0105	0.0167	0.0186	0.0815	0.0349
SSD-Resnet & Dlib ResNet & NB	0.0078	0.5956	0.0182	0.7255	0.0105	0.7009	0.0246	0.6904	0.2934	0.7367
SSD-Resnet & Dlib ResNet & Linear SVM	0.0070	0.3656	0.0075	0.4888	0.0017	0.5798	0.0016	0.5904	0.0379	0.7079
SSD-Resnet & Dlib ResNet & RBF SVM	0.0030	0.3365	0.0044	0.4762	0.0019	0.5492	0.0016	0.5561	0.0301	0.6770
SSD-Resnet & Dlib ResNet & k-NN	0.7746	0.7242	0.7930	0.7787	0.8086	0.7870	0.8003	0.7500	0.8147	0.8047
SSD-Resnet & Dlib ResNet & DT	0.0082	0.0100	0.0320	0.0270	0.0496	0.0227	0.0740	0.0596	0.1369	0.3461
SSD-Resnet & Dlib ResNet & RF	0.0697	0.1069	0.0845	0.2354	0.0981	0.3073	0.1429	0.2497	0.3247	0.4826
SSD-Resnet & Dlib ResNet & NN	0.0377	0.1886	0.0296	0.5148	0.0300	0.7080	0.0221	0.7538	0.2934	0.7926
SSD-Resnet & Dlib ResNet & AB	0.0199	0.0633	0.0304	0.1645	0.0508	0.1885	0.1006	0.1875	0.2424	0.3548
SSD-Resnet & Dlib ResNet & QDA	0.0199	0.0173	0.0304	0.0175	0.0508	0.0366	0.1006	0.0356	0.2424	0.3034
SSD-Resnet & FaceNet & NB	0.0052	0.0362	0.0245	0.0469	0.0303	0.0468	0.0397	0.0737	0.0990	0.1129
SSD-Resnet & FaceNet & Linear SVM	0.0052	0.0362	0.0245	0.0469	0.0303	0.0468	0.0010	0.0737	0.0990	0.1129
SSD-Resnet & FaceNet & RBF SVM	0.0052	0.0362	0.0245	0.0469	0.0303	0.0468	0.0026	0.0737	0.0990	0.1129
SSD-Resnet & FaceNet & k-NN	0.0052	0.0362	0.0245	0.0469	0.0303	0.0468	0.4577	0.0737	0.0990	0.1129
SSD-Resnet & FaceNet & DT	0.0052	0.0362	0.0245	0.0469	0.0303	0.0468	0.0372	0.0737	0.0990	0.1129
SSD-Resnet & FaceNet & RF	0.0052	0.0362	0.0245	0.0469	0.0303	0.0468	0.0942	0.0737	0.0990	0.1129
SSD-Resnet & FaceNet & NN	0.0052	0.0362	0.0245	0.0469	0.0303	0.0468	0.1135	0.0737	0.0990	0.1129
SSD-Resnet & FaceNet & AB	0.0052	0.0362	0.0245	0.0469	0.0303	0.0468	0.0679	0.0737	0.0990	0.1129
SSD-Resnet & FaceNet & QDA	0.0052	0.0362	0.0245	0.0469	0.0303	0.0468	0.0679	0.0737	0.0990	0.1129
MTCNN & OpenFace & NB	0.0050	0.0028	0.0142	0.0108	0.0076	0.0116	0.0142	0.0184	0.0271	0.0302
MTCNN & OpenFace & Linear SVM	0.0000	0.0028	0.0023	0.0108	0.0007	0.0116	0.0142	0.0184	0.0041	0.0302
MTCNN & OpenFace & RBF SVM	0.0005	0.0028	0.0043	0.0108	0.0033	0.0116	0.0142	0.0184	0.0072	0.0302
MTCNN & Open Face & k-NN	0.1143	0.0028	0.1670	0.0108	0.1700	0.0116	0.0142	0.0184	0.3183	0.0302
MTCNN & OpenFace & DT	0.0064	0.0028	0.0117	0.0108	0.0189	0.0116	0.0142	0.0184	0.0450	0.0302
MTCNN & OpenFace & RF	0.0434	0.0028	0.0478	0.0108	0.0832	0.0116	0.0142	0.0184	0.1372	0.0302
MTCNN & OpenFace & NN	0.0703	0.0028	0.0774	0.0108	0.1107	0.0116	0.0142	0.0184	0.0415	0.0302

Models	143 classes		85 classes		57 classes		41 classes		19 classes	
	Not augmented	Augmented	Not augmented	Augmented	Not augmented	Augmented	Not augmented	Augmented	Not augmented	Augmented
MTCNN & OpenFace & AB	0.0127	0.0028	0.0489	0.0108	0.0315	0.0116	0.0142	0.0184	0.1008	0.0302
MTCNN & OpenFace & QDA	0.0127	0.0028	0.0489	0.0108	0.0315	0.0116	0.0142	0.0184	0.1008	0.0302
MTCNN & Dlib ResNet & NB	0.0059	0.3817	0.0188	0.4835	0.0080	0.5502	0.0169	0.4462	0.0271	0.6095
MTCNN & Dlib ResNet & Linear SVM	0.0000	0.7512	0.0009	0.8595	0.0007	0.8581	0.0000	0.8639	0.0041	0.9012
MTCNN & Dlib ResNet & RBF SVM	0.0002	0.6832	0.0003	0.7924	0.0007	0.8081	0.0000	0.8185	0.0020	0.8654
MTCNN & Dlib ResNet & k-NN	0.8123	0.8239	0.8857	0.8885	0.8910	0.8943	0.8886	0.8598	0.9217	0.9094
MTCNN & Dlib ResNet & DT	0.0193	0.0083	0.0739	0.0580	0.0365	0.0404	0.0408	0.0600	0.0619	0.2462
MTCNN & Dlib ResNet & RF	0.0772	0.1499	0.1487	0.2506	0.2310	0.3338	0.1856	0.3930	0.2917	0.5896
MTCNN & Dlib ResNet & NN	0.0482	0.4455	0.0256	0.5762	0.0172	0.8101	0.0169	0.8766	0.0271	0.8884
MTCNN & Dlib ResNet & AB	0.0198	0.0080	0.0538	0.0535	0.1028	0.0345	0.0964	0.1057	0.1029	0.3306
MTCNN & Dlib ResNet & QDA	0.0198	0.0080	0.0538	0.0100	0.1028	0.0361	0.0964	0.0405	0.1029	0.3076
MTCNN & FaceNet & NB	0.0042	0.0372	0.0111	0.0404	0.0076	0.0431	0.0664	0.0589	0.0271	0.1269
FaceNet & OpenFace & Linear SVM	0.0000	0.0630	0.0023	0.1263	0.0007	0.1299	0.0041	0.1320	0.0031	0.1270
FaceNet & OpenFace & RBF SVM	0.0005	0.0574	0.0043	0.0904	0.0017	0.1385	0.0124	0.1207	0.0077	0.3367
FaceNet & Open Face & k-NN	0.1164	0.0659	0.1638	0.1297	0.1720	0.1177	0.1729	0.1466	0.3321	0.2124
FaceNet & OpenFace & DT	0.0054	0.0106	0.0225	0.0102	0.0199	0.0182	0.0326	0.0442	0.0461	0.0885
FaceNet & OpenFace & RF	0.0165	0.0305	0.0407	0.0475	0.0673	0.0726	0.0645	0.0716	0.1080	0.1837
FaceNet & OpenFace & NN	0.0710	0.0531	0.0757	0.1024	0.1223	0.1127	0.0960	0.1125	0.0328	0.2646
FaceNet & OpenFace & AB	0.0127	0.0090	0.0344	0.0097	0.0421	0.0133	0.0585	0.0337	0.1039	0.0415
FaceNet & OpenFace & QDA	0.0127	0.0068	0.0344	0.0108	0.0421	0.0149	0.0585	0.0277	0.1039	0.2984
FaceNet & Dlib ResNet & NB	0.0084	0.3834	0.0108	0.4849	0.0080	0.5406	0.0169	0.4421	0.0271	0.6024
FaceNet & Dlib ResNet & Linear SVM	0.0002	0.7476	0.0006	0.8544	0.0007	0.8598	0.0000	0.8616	0.0036	0.9017
FaceNet & Dlib ResNet & RBF SVM	0.0002	0.6759	0.0003	0.7958	0.0007	0.8127	0.0000	0.7945	0.0020	0.8634
FaceNet & Dlib ResNet & k-NN	0.8135	0.8239	0.8842	0.8874	0.8919	0.8943	0.8886	0.8594	0.9232	0.9084
FaceNet & Dlib ResNet & DT	0.0189	0.0116	0.0304	0.0284	0.0779	0.0222	0.0255	0.0664	0.0686	0.2344
FaceNet & Dlib ResNet & RF	0.0725	0.1551	0.1493	0.3049	0.1511	0.2883	0.1995	0.3067	0.4734	0.5328
FaceNet & Dlib ResNet & NN	0.0541	0.4585	0.0262	0.5552	0.0080	0.8121	0.0169	0.8733	0.0271	0.8884
FaceNet & Dlib ResNet & AB	0.0118	0.0102	0.0401	0.0427	0.0776	0.0663	0.0855	0.1219	0.1039	0.3347
FaceNet & Dlib ResNet & QDA	0.0118	0.0102	0.0401	0.0100	0.0776	0.0365	0.0855	0.0397	0.1039	0.3086
FaceNet & FaceNet & NB	0.0047	0.0626	0.0105	0.0392	0.0080	0.0722	0.0169	0.0769	0.0271	0.1807
FaceNet & FaceNet & Linear SVM	0.0005	0.2875	0.0014	0.2844	0.0013	0.3271	0.0015	0.2767	0.0031	0.3536
FaceNet & FaceNet & RBF SVM	0.0032	0.0918	0.0017	0.0518	0.0000	0.0679	0.0049	0.1053	0.0241	0.1162
FaceNet & FaceNet & k-NN	0.3021	0.3762	0.3842	0.4337	0.4461	0.4713	0.4942	0.3547	0.5415	0.5061
FaceNet & FaceNet & DT	0.0123	0.0094	0.0176	0.0162	0.0182	0.0169	0.0176	0.0199	0.1325	0.0425
FaceNet & FaceNet & RF	0.0375	0.0737	0.0566	0.0904	0.0965	0.1173	0.0874	0.1324	0.1100	0.2958
FaceNet & FaceNet & NN	0.1803	0.3019	0.2392	0.3706	0.1936	0.4309	0.1616	0.3135	0.1566	0.4749
FaceNet & FaceNet & AB	0.0177	0.0113	0.0185	0.0108	0.0633	0.0129	0.0536	0.0229	0.0502	0.3117
FaceNet & FaceNet & QDA	0.0177	0.0380	0.0185	0.0364	0.0633	0.0663	0.0536	0.0739	0.0502	0.1274

Acknowledgment

This work was supported by the Spanish Agencia Estatal de Investigación. Project Monitoring and tracking systems for the improvement of intelligent mobility and behavior analysis (SiMoMIAC). PID2019-108883RB-C21 / AEI /

10.13039/501100011033. The research of Diego M. Jiménez-Bravo has been co-financed by the European NextGenrationEU Fund, Spanish "Plan de Recuperación, Transformación y Resilencia" Fund, Spaninsh Ministry of Universities and University of Salamanca ("Ayudas para la recualificación del sistema universitario español para 2021-2023"). André Filipe Sales Mendes's research was co-financed by the European Social Fund and Junta de Castilla y León (Operational Programme 2014-2020 for Castilla y León, EDU/556/2019 BOCYL).

References

[1] E. G. Miller, N. E. Matsakis, P. A. Viola, "Learning from one example through shared densities on transforms," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 464– 471, 2000, doi: 10.1109/CVPR.2000.855856.

[2] L. Fei-Fei, R. Fergus, P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 594–611, 4 2006, doi: 10.1109/TPAMI.2006.79.

[3] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, 6 2016, pp. 3637–3645, Neural information processing systems foundation.

[4] A. Shaban, S. Bansal, Z. Liu, I. Essa, B. Boots, "One-shot learning for semantic segmentation," *British Machine Vision Conference 2017, BMVC 2017*, 9 2017, doi: 10.5244/c.31.167.

[5] M. Woodward, C. Finn, "Active One-shot Learning," *arXiv*, 2 2017.

[6] P. Wang, L. Liu, C. Shen, Z. Huang, A. Van Den Hengel, H. T. Shen, "Multi-attention network for one shot learning," 2017. doi: 10.1109/CVPR.2017.658.

[7] H. Altae-Tran, B. Ramsundar, A. S. Pappu, V. Pande, "Low Data Drug Discovery with One-Shot Learning," *ACS Central Science*, vol. 3, pp. 283–293, 4 2017, doi: 10.1021/acscentsci.6b00367.

[8] C. Shorten, T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, pp. 1–48, 12 2019, doi: 10.1186/s40537-019-0197-0.

[9] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, "Random erasing data augmentation," 8 2020. [Online]. Available: https://github.com/zhunzhong07/Random-Erasing., doi: 10.1609/aaai.v34i07.7000.

[10] L. Perez, J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," *arXiv*, 12 2017.

[11] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, pp. 2613–2617, 4 2019, doi: 10.21437/Interspeech.2019-2680.

[12] A. Kumar, A. Kaur, M. Kumar, "Face detection techniques: a review," *Artificial Intelligence Review*, vol. 52, pp. 927–948, 8 2019, doi: 10.1007/s10462-018- 9650-2.

[13] P. Viola, M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 5 2004, doi: 10.1023/B:VISI.0000013087.49260.fb.

[14] A. Jadhav, S. Lone, S. Matey, T. Madamwar, S. Jakhete, "Survey on Face Detection Algorithms," 2021. [Online]. Available: www.ijisrt.com.

[15] X. Lu, X. Kang, S. Nishide, F. Ren, "Object detection based on SSD-ResNet," in *Proceedings of 2019 6th IEEE International Conference on Cloud Computing and Intelligence Systems, CCIS 2019*, 12 2019, pp. 89–92, Institute of Electrical and Electronics Engineers Inc.

[16] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, pp. 1499–1503, 4 2016, doi: 10.1109/LSP.2016.2603342.

[17] Chunming Wu, Ying Zhang, "MTCNN and FACENET Based Access Control System for Face Detection and Recognition," *Automatic Control and Computer Sciences*, vol. 55, pp. 102–112, 1 2021, doi: 10.3103/S0146411621010090.

[18] S. Milborrow, F. Nicolls, "Locating facial features with an extended active shape model," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5305 LNCS, 10 2008, pp. 504–513, Springer Verlag.

[19] J. Saragih, R. Goecke, "A nonlinear discriminative approach to AAM fitting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2007.

[20] N. Kumar, P. Belhumeur, S. Nayar, "FaceTracer: A search engine for large collections of images with faces," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5305 LNCS, 10 2008, pp. 340–353, Springer Verlag.

[21] X. Cao, Y. Wei, F. Wen, J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014, doi: 10.1007/s11263-013-0667-3.

[22] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, S. Zafeiriou, "Mnemonic Descent Method: A Recurrent Process Applied for End-to-End Face Alignment," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016- Decem, 12 2016, pp. 4177–4187, IEEE Computer Society.

[23] J. Lv, X. Shao, J. Xing, C. Cheng, X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," 2017. doi: 10.1109/CVPR.2017.393.

[24] J. Yang, Q. Liu, K. Zhang, "Stacked Hourglass Network for Robust Facial Landmark Localisation," 2017. [Online]. Available: https://www.umdfaces.io., doi: 10.1109/CVPRW.2017.253.

[25] A. Newell, K. Yang, J. Deng, "Stacked hourglass networks for human pose estimation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9912 LNCS, pp. 483–499, 3 2016, doi: 10.1007/978- 3-319-46484-8_29.

[26] J. Deng, G. Trigeorgis, Y. Zhou, S. Zafeiriou, "Joint Multi-View Face Alignment in the Wild," *IEEE Transactions on Image Processing*, vol. 28, pp. 3636–3648, 8 2019, doi: 10.1109/TIP.2019.2899267.

[27] A. Ahmed, J. Guo, F. Ali, F. Deeba, A. Ahmed, "LBPH based improved face recognition at low resolution," in *2018 International Conference on Artificial Intelligence and Big Data, ICAIBD 2018*, 6 2018, pp. 144–147, Institute of Electrical and Electronics Engineers Inc.

[28] T. Baltrusaitis, A. Zadeh, Y. C. Lim, L. P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, 6 2018, pp. 59–66, Institute of Electrical and Electronics Engineers Inc.

[29] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, 12 2016, pp. 770– 778, IEEE Computer Society.

[30] F. Schroff, D. Kalenichenko, J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 815–823, 3 2015, doi: 10.1109/CVPR.2015.7298682.

[31] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, H. Chen, Y. Miao, Z. Huang, J. Liang, "Masked Face Recognition Dataset and Application," 3 2020.

[32] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," 2018. doi: 10.1109/CVPR.2018.00552.

[33] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, "SphereFace: Deep hypersphere embedding for face recognition," 2017. doi: 10.1109/CVPR.2017.713.

[34] A. Suruliandi, A. Kasthuri, S. P. Raja, "Deep Feature Representation and Similarity Matrix based Noise Label Refinement Method for Efficient Face Annotation," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, pp. 2–66, doi: 10.9781/ijimai.2021.05.001.

[35] "European Edge Computing Consortium(EECC)." [Online]. Available: https://ecconsortium.eu/.

[36] X. Jiang, F. R. Yu, T. Song, V. C. Leung, "Intelligent Resource Allocation for Video Analytics in Blockchain- Enabled Internet of Autonomous Vehicles with Edge Computing," *IEEE Internet of Things Journal*, pp. 1–1, 9 2020, doi: 10.1109/jiot.2020.3026354.

[37] M. Gheisari, G. Wang, S. Chen, "An Edge Computing- enhanced Internet of Things Framework for Privacy-preserving in Smart City," *Computers and Electrical Engineering*, vol. 81, p. 106504, 1 2020, doi: 10.1016/j.compeleceng.2019.106504.

[38] H. Albataineh, M. Nijim, D. Bollampall, "The Design of a Novel Smart Home Control System using Smart Grid Based on Edge and Cloud Computing," *2020 8th International Conference on Smart Energy Grid Engineering, SEGE 2020*, pp. 88–91, 8 2020, doi: 10.1109/SEGE49949.2020.9181961.

[39] K. P. Dirgantoro, J. M. Lee, D. S. Kim, "Generative Adversarial Networks Based on Edge Computing with Blockchain Architecture for Security System," *2020 International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2020*, pp. 039– 042, 2 2020, doi: 10.1109/ICAIIC48513.2020.9065231.

[40] L. J. Karam, T. Zhu, "Quality labeled faces in the wild (QLFW): a database for studying face recognition in real-world environments," 2015. [Online]. Available: http://vis-www.cs.umass.edu/lfw/., doi: 10.1117/12.2080393.

Diego M. Jiménez-Bravo

He studied a degree in Computer Engineering Management and Information Systems at the University of the Basque Country (UPV/EHU) (2016). Subsequently, he obtained a Master's degree in Intelligent Systems from the University of Salamanca (USAL) (2017). He concluded his formative stage by obtaining a Ph.D. in Computer Engineering from the USAL (2020). He is part of the research group ESALab, of the same entity, where he carries out research work under a postdoctoral contract funded by the European NextGenrationEU Fund, Spanish "Plan de Recuperación, Transformación y Resilencia" Fund, Spanish Ministry of Universities, and the University of Salamanca. His main research interests focus on the field of artificial intelligence, IoT (Internet of Things), smart-homes, energy optimisation and social networks, among others. He also collaborates in several research projects within the research group.

Álvaro Lozano Murciego

He received the master's degree in intelligent systems and the Ph.D. degree in computer engineering from the University of Salamanca, in 2015 and 2019, respectively. He is currently working on the Expert Systems and Applications Laboratory Research Group, Computer Science Department, University of Salamanca, as an Assistant Professor. Throughout his training, he has followed a well-defined line of research, focused on machine learning, route optimization, IoT sensors, and edge computing.

André Sales Mendes

He is a Ph.D. student in computer engineering. He has completed a computer engineering degree and a master's degree in intelligent systems at the University of Salamanca. He is currently focusing his doctoral studies on artificial intelligence and expert robots. He is a member of the Expert Systems And Applications Laboratory research group at the University of Salamanca. In addition, he has several publications in JCR indexed journals.

Luis Augusto Silva

He received the degree in Internet systems from the Federal Institute of Santa Catarina (IFC), Camboriú, Brazil, in February 2017, and the master's degree in applied computing from the University of Itajaí Valley, Brazil, in 2019. He is currently pursuing the Ph.D. degree incomputer engineering with the University of Salamanca, Spain. His research during his master's degree included the field of notification systems,the IoT, and data privacy. During the master's degree, he was a collaborating Researcher with the Laboratory of Embedded and Distributed Systems (LEDS), UNIVALI, collaborating with research projects related to the Internet of Things. Since August 2020, he has been a Researcher with the Expert Systems and Applications Laboratory (ESALab), Universidad de Salamanca. His research interests include the Internet of Things, embedded drone systems, and data privacy applied to smart environments.

Daniel H. De La Iglesia

He is a Ph.D. in Computer Engineering, a Technical Engineer in Computer Systems and a Degree in Computer Engineering from the University of Salamanca. He has an official master's degree in Intelligent Systems and in recent years has been linked to different research groups in the field of artificial intelligence where he has participated in dozens of national and international research projects. He is the author of different book chapters and has presented more than twenty research papers in different international congresses. In addition, it has numerous scientific publications in international impact journals indexed in the JCR reference ranking. He has been awarded the first prize of the Open Data Contest organized by the Junta de Castilla y León (Spain) in 2013, and with the first innovation award for the best research project awarded by the Junta de Castilla y León (Spain) in 2016. He is currently a Professor at the Pontifical University of Salamanca.

Promoting Social Media Dissemination of Digital Images Through CBR-Based Tag Recommendation

Lucía Martín-Gómez*, Javier Pérez-Marcos, Rebeca Cordero-Gutiérrez, Daniel H. de la Iglesia

Department of Computer Science, Pontifical Univeristy of Salamanca C/Compañía, 5, 37002 Salamanca (Spain)

Received 15 April 2022 | Accepted 29 June 2022 | Early Access 16 September 2022

Abstract

Multimedia content has become an essential tool to share knowledge, sell products or disseminate messages. Some social networks use multimedia content to promote information and create social communities. In order to increase the impact of the digital content, those images or videos are labeled with different words, denominated tags. In this paper, we propose a recommender system which analyzes multimedia content and suggests tags to maximize its influence in the social community. It implements a Case-Based Reasoning architecture (CBR), which allows to learn from previous tagged content. The system has been evaluated through cross fold validation with a training and validation sets carefully constructed and extracted from Instagram. The results demonstrate that the system can suggest good options to label our image and maximize the influence of the multimedia content.

Keywords

Artificial Intelligence, Digital Image Processing, Recommender System, Social Media, Tagging.

DOI: 10.9781/ijimai.2022.09.002

I. Introduction

The information is the essence of any communication support. The data offered to the individuals provokes a reaction to consume them, either because of its quality, its originality or the way in which it is told. In the particular case of Internet, this issue becomes fundamental for the knowledge sharing. With regard to a digital entity (i.e. a webpage, a social profile or a digital product) it is difficult to achieve the objectives for which it is created when interesting information for its visitors is not present or it is poorly ranked.

As we live in a multimedia world, the information has stopped being limited to a text, but to a mixture of digital objects that allow us to transmit a message. Therefore, text is supported by other visual elements, denominated multimedia content, that serve to draw the attention of our audience. Thus, multimedia content becomes a more attractive alternative for those users who prefer this type of support instead of reading on the website. In this way, the information is mainly based on images and conveys the message we are trying to promote. Furthermore, the continuous increase of users connected to the Web requires new methods to maximize the impact of information dissemination.

In particular, social networks have taken advantage of the power of multimedia content to promote their data. Moreover, some social networks such as Instagram or TikTok have made the multimedia content their essence to survive and expand in the Internet.

Instagram[1] is a social network to upload photos and videos. The users can also apply photographic effects such as filters or frames, add text, gifs and stickers to their posts or create compilations of several short video fragments. Despite its recent birth in 2010, its concept has been rapidly accepted by the society, and by the end of 2021 it had more than 2 billion active users [1].

Images in many social networks are promoted through the use of tags, which consist of short words that somehow describe the content or the purpose of the picture. Tags are essential for the correct dissemination of multimedia content through the social platform. The tagging of multimedia content to categorize publications by subject matter on the social media is done through the so-called hashtags, so henceforth tags or hashtags will be referred to indistinctly. There are a series of metrics that are calculated based on the interactions of other users with the post (like, share the post, write a comment, save the post...) Thus, the selection of the words to tag the multimedia content becomes essential to augment the visibility of the image and therefore the user. However, despite of different proposals to tag different kind of content [2]–[4] there is no standard method in social networks to know beforehand which words are better to optimize the impact of the image.

In this paper, we present a recommender system that suggests tags to promote a digital image submitted to social networks. In order to improve the performance, the recommender consists of a Case-Based Reasoning (CBR) architecture, which is able to learn from previous experiences to obtain better results in the future. Initially, the memory of the system is previously populated with image features obtained from a set of photos uploaded to Instagram and their associated tags. Then, the system can recommend tags for a particular image manually selected.

For this purpose, the main features of the image are extracted and analyzed. With these features, we obtain a map which is compared with previous images stored in the memory and selected those ones which are applicable due to their similarity. Finally, a set of words are

[1] https://www.instagram.com/

* Corresponding author.
E-mail address: lmartingo@upsa.es

chosen following a set of rules created. Since this process coincides with the theoretical approach of a CBR and taking into account that the literature shows that this type of systems obtain very good results in tagging and recommendation problems, our proposal implements a CBR for the image tagging task.

The experiments aim to demonstrate the performance of the system independently of the dataset, by applying a cross validation. Additionally, we aim to prove that CBR can suggest good tags to label multimedia content. For this purpose, a comparison between the CBR system and other regular classification systems was performed.

The remainder of the paper is structured as follows. Section II briefly describes recent works related to the recommendation in multimedia content. Section III discusses the techniques used for image processing. Section IV provides the technical details of the recommender system. Sections V and VI detail the case study in Instagram and experiments carried out to validate this proposal and discusses the preliminary results. Finally, Section VII exposes the main contributions of this work.

II. Recommender Systems for Multimedia Content

Feature extraction of multimedia content has been deeply explored to create recommender systems. In music, [5] applies a set of boosted classifiers to map audio features onto social tags collected from the Web. The resulting automatic tags are part of a social recommender. [6] predicts potentially interesting and unknown music based on an analysis of musical features of musical tracks. [7] proposes a recommender system to suggest music by applying data mining techniques with information about its content and the context. [8] proposes a model for recommendation to predict the latent factors from music audio when they cannot be obtained from usage data. [9] learns features from audio content and makes personalized recommendations. In images, [10] presents an analyzer to extract features from images for recommendation purposes. [11] proposes a progressive image search and recommendation system, which incorporates the auto-interpretation and user behavior.

There are some approaches that suggest new tags for specific digital content. For instance, [2] present TagAssist, a recommender system that recommend tags for posts by applying a Case Based Reasoning (CBR) architecture. [12] propose a strategy that enables a content-based recommender to infer user interests by applying machine learning techniques both on the "official" item descriptions provided by a publisher, and on tags which users adopt to annotate relevant items. More recently, [3] proposes a new framework which makes recommendation of tag-based multimedia recipe. [4] framework that is able to utilize knowledge over the Linked Open Data (LOD) cloud to recommend context-based services to users. However, these proposals do not consider the multimedia content for the calculation of new tags.

Additionally, recommender systems commonly makes use of learning techniques, specially CBR, when multimedia content is involved. [13] makes use of a CBR to tag emotions from facial expressions. [14] presents a new recommender with a CBR that exploits audio and tagging knowledge using a hybrid representation and adding semantic knowledge extracted from the tags of similar music tracks. [15] presents a medical CBR system with a knowledge-based recommendation, which analyzes image and text from patient health records. [16] presents a CBR that exploits nuclear image features to retrieve the cases that are the most similar to the new image test and to compute the most probable diagnoses. [17] develops a CBR System for face recognition under partial occlusion. All the proposals obtained very successful results, even when compared with other similar techniques used for the same purposes.

In terms of automatic labeling of multimedia information, the results obtained are not fully satisfactory, so most proposals opt for human-machine collaboration for more accurate and efficient multimedia tagging [18]. [19] proposes a D2-clustering-based method that represents the multimedia content by bags of weighted vectors. On the other hand, [20] describes a scalable algorithm that considers the use of matching labels in images with similar characteristics by accumulating votes from visually similar neighbors. Other proposals, such as [21], combine historical image tagging information and metadata into an adaptive factorization model that applies transfer learning and deep learning image classification techniques.

Another important concept in tagging and recommendation systems for multimedia content is that of folksonomies, which is a system for assigning tags to elements by users [22]. Thus, the assignment of a tag for a particular content is influenced by the general criteria of the users [23]. Some authors have already made use of folksonomies in multimedia content tagging problems. When social tagging generates folksonomies in image-related content, these are called visual folksonomies. For instance, [24] describes some techniques for automatic image tagging that take benefit of collaboratively tagged image databases and [25] formulates image tag recommendation as a maximum a posteriori (MAP) problem, making use of a visual folksonomy.

As it is demonstrated with some recent related work, CBR can get very good results in tagging and recommender systems [26], [27]. This careful literature review leads us to build a new CBR-based assistive tagging system capable of analyzing multimedia content and suggesting tags to maximize its impact on social networks.

III. Image Feature Extraction

As we stated in the introduction, the present work aims to recommend tags in order to promote multimedia content in a social network. Recommenders usually retrieve information from different data provided by the user, such as personal information or navigation data. Companies Amazon, Spotify or TripAdvisor find that information as essential for the correct operation of their recommenders [28].

It becomes clear that the feature extraction and its analysis is needed for the recommendation process. Therefore, starting point of this system is the extraction of the main features of a digital image. At this step, if the multimedia content submitted to the social network is a video file, its representative image (video thumbnail) is considered. Then, the extracted visual descriptors are compared with those of other images' to search for similar experiences.

An image is described by the shape and layout of its elements, as well as its colors. These types of descriptors are the main features considered in image preprocessing tasks. Due to their nature, these parameters are usually divided into two categories: shape and disposition, and color. This taxonomy permits to address different problems in which not every feature of the image should be involved. As an example we can cite the identification of a particular object in an image, which does not consider the particular colors, or the search of color histogram, which does not depend on the shape and disposition of figures in a picture.

In this sense, there are different techniques to detect the shape and disposition of the elements that are part of an image. For this particular problem, we focus the algorithms that capture invariant image descriptors, as we aim to detect similarities even if the image is rotated or transformed. Among the different proposals, Scale-Invariant Feature Transform (SIFT) [29] is of particular interest due to their successful results in different problems. Some versions of SIFT, such as Speeded Up Robust Features (SURF) [30] and Oriented FAST and

Rotated BRIEF (ORB) [31], were proposed to improve the computation time of the original algorithm, which in some cases, can become a little high. In order to also reduce the dimensionality of the problem and improve the times by grouping the descriptors into clusters, Bag of Visual Words (BoVW) [32] is frequently applied.

To analyze the color of a digital image, color histograms have been long-established in this field [33]. This technique obtains the color distribution based on the values of the pixels. However, the immense amount of pixels, and therefore, colors that compose a digital image involves a great difficulty in the color extraction process. As a solution, color quantization procedure is widely used to reduce the number of colors in an image, extracting only the most representative ones [34]. The color quantization can be implemented following different algorithms, such as clustering, k-means or neural networks. It is essential to analyze each proposal and select which one can better adapt to solve our problem.

Moreover, transfer learning involves the exploitation of learning outcomes to a related task [35]. In neural networks, transfer learning is applied by obtaining the weights of the layers prior to the classification of a model already trained for a similar problem, which are given as feature vectors called embeddings. [36] applies this technique with the Inception-v3 model [37] to classify different Instagram influencer profiles according to their interests or posting topics (i.e. fashion or beauty). Thus, this technique reduces the computational cost of feature extraction, optimizes the results by working with already validated data related to the problem to be treated and makes the feature extraction phase much more efficient [38].

After a careful examination of the literature and taking into account the main objective of this work in which the execution time is not critical, an original version of the SIFT has been considered. In order to reduce the dimensionality of the problem, BoVW is also applied in the image feature extraction process. For the color extraction, a color histogram was obtained to represent chromatic information from the images. Additionally, in this proposal, transfer learning will also be applied based on the embeddings obtained from the Inception-v3 model [37].

IV. Recommender System

This paper proposes a CBR-based recommender system that relies on image metadata to propose tags for disseminating multimedia content. Thus, the image is the input to a system that generates, as a recommendation, a set of tags appropriate to its content.

Our proposal is based on the CBR methodology combined with state-of-the-art techniques within the field of deep learning. Unlike other content-based recommender systems, this work uses Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN) in two different ways: on the one hand they are used to infer a Case-Representation feature space, and on the other hand to define a hashtag latent space embedding. The definition of a Case-Representation feature space allows the CBR stages of retrieval and reuse to use a distance-based recommendation. The creation of a latent space of hashtags allows that given a hashtag to recommend we can obtain those closest ones that are more likely to be used in combination.

The Case-Base of our CBR is formed by Case-Representations obtained from an initial set of images labeled with hashtags. This set of images will be used to train the Case Representation Neural Network and subsequently to initialize the Case-Base, as shown in Fig. 1. The idea of using a Case Representation Neural Network (CRNN) is to obtain a Case-Representation space where the items on which the same tags are used are close and at the same time far away from those that do not use them. In this way, we will be able to use distances in the retrieval phase, such as the squared distance, on the cases to obtain the closest ones [17]. The CRNN is formed by the embedding layer of a Siamese Neural Network using Triplet Loss [39].

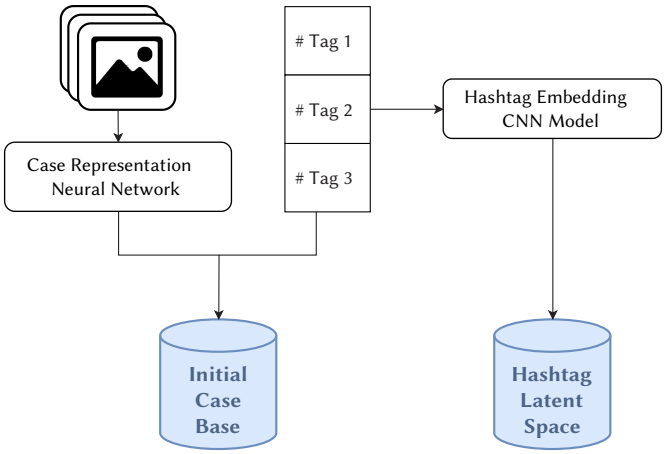


Fig. 1. Overview of the case-base inizialization and hashtag latent space creation process.

The hashtags latent space is formed by the Neural Network Embeddings of the hashtags. This initial latent space is obtained from the same CRNN training set as shown in Fig. 1, but in this case using the hashtags. When recommending hashtags, the reuse phase of the CBR will propose the hashtag that best suits the new case. However, more than one hashtag per image must be recommended. To this end, once the proposed solution has been obtained, the latent space of hashtags will be searched for those that come closest to the proposal. To obtain this latent space of hashtags, a DNN has been trained to obtain the embeddings, so that those hashtags that are often used together are close in this latent space, while those that are never used together are far away.

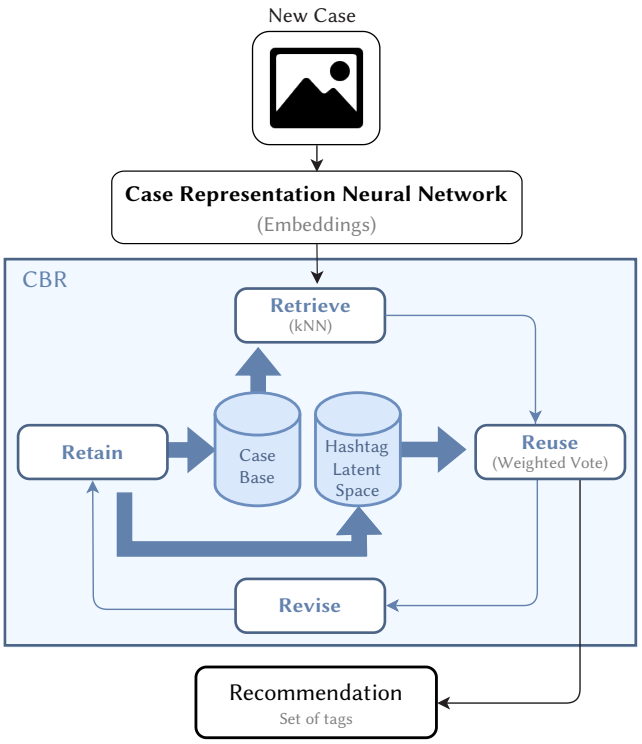


Fig. 2. Overview of the proposed CBR-base recommender system.

In order to be able to make recommendations, the Case-Base must have been initialized and the latent space of hashtags must have been defined. To make a recommendation on a new image, its Case-Representation must first be obtained. To do this, the image is passed through the CRNN to obtain its embedding. In the retrieve phase, the most relevant cases of the Case-Base are obtained from the Case-Representation of the new case and by means of the squared distance. In the following reuse phase, the proposed hashtag is obtained by a weighted vote of the retrieved cases. From the obtained proposal, the k closest hashtags are searched in the latent space of hashtags, forming together the recommendation. In the revision phase, we check which of the hashtags finally used were not recommended, to store the Case-Representation together with the non-recommended hashtags in the retention phase. These last two phases allow our system to be able to adapt to new hashtags and learn from users' tagging habits. The complete cycle of the proposed CBR can be seen in Fig. 2.

A. *Retrieve: Getting the Best Tags*

In the retrieval phase of a CBR, the system recovers from the Case-Base the cases most similar to the Case-Representation of the new case. The Case-Representation of our proposed system for an image x is an embedding $f(x)$ such that in a feature space R^d the squared distance of identically labeled images is small, while for differently labeled images it is large.

1. Image Embedding Network Architecture

The CNN architecture responsible for the image embedding can be seen in Fig. 3. The first stages of the network are reused from another pre-trained network for the task of classifying images. This technique is known as transfer learning. In this way we can obtain a feature vector from an image, without the need to retrain it. The pre-trained network for our proposal is Inception-v3 due to its outstanding performance [37]. In order to use a pre-trained network, the final stages in charge of classification must be removed. In the case of Inception-v3 we have removed the last two layers (an Average-Pooling pre-classification layer and a fully-connected layer), leaving a final layer of (8 x 8 x 2048) components.

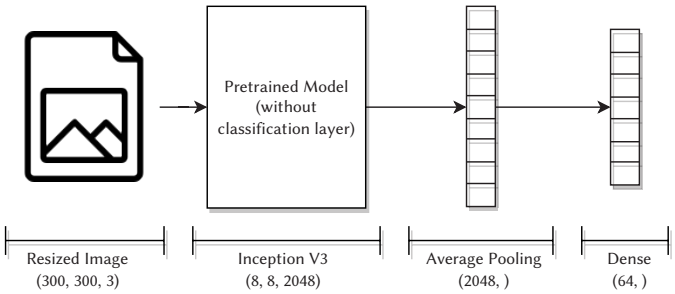


Fig. 3. Overview of the Case Representation Neural Network architecture.

The next two layers added to the modified pre-trained network are in charge of obtaining the embeddings of the images. The first one consists of a 1-dimensional Average-Pooling layer. This layer allows to reduce the dimensionality of the feature maps of the previous layer, making it more robust to changes in the positions of the image features. The second layer consists of a Dense layer, using the hyperbolic tangent as an activation function. The output of this layer will be the embedding of the image, so the size of the layer will determine the size of the embedding. In our case we have chosen a size of 64 components for this layer.

2. Siamese Network With Triplet Loss Architecture

As mentioned above, a Case-Representation has to be a set of features from which we can at the retrieval stage recover those cases of the Case-Base that are most similar. It is important that the set of features forming the Case-Representation represents an image in an embedded space in such a way that semantically related images are metrically close. For this purpose, a Siamese Neural Network has been used together with the Triplet Loss as a cost function.

The Siamese Neural Network structure consists of two branches formed by the same neural network model that share the weights and parameters [40]. During training, the network is fed with image pairs. The objective of this network is to learn the optimal features of the images in such a way that related images are pulled closer while those that are not pushed away. To optimize the neural network, a cost function capable of making a distinguish between pair is used defined in (1).

$$\mathcal{L} = \frac{1}{2} l D^2 + \frac{1}{2}(1-l)\{\mathrm{m}\ \ (0, m-D)\}^2 \tag{1}$$

Where l is a binary label selecting whether the input pair consisting of image x_1 and x_2 is a positive ($l = 1$) or negative ($l = 0$), m>0 is the margin for dissimilar pairs and $D = \|f(x_1) - f(x_2)\|_2$ is the Euclidean distance between feature vectors $f(x_1)$ and $f(x_2)$ of input images x_1 and x_2.

The neural network used in the proposed system is a variation of the Siamese Neural Network called Triplet Neural Network [39]. Unlike the Siamese Neural Network, this one consists of three branches with the same neural network model, sharing the same weights and features. The input of this network is formed by a triplet of objects. While in a Siamese Neural Network the pairs of objects could be related or unrelated, in a Triplet Neural Network one of the objects is the anchor, while of the remaining two one is related to the anchor (positive) and the other is unrelated (negative). Formally, for the triplet (x^a, x^p, x^n) one (x^a is the anchor, x^p is the positive and x^a is the negative) has that r(x^a, x^p)>r (x^a, x^n) where $r(.)$ is a similarity measure. The cost function of the Triplet Neural Network is the Triplet Loss Function. We want the image x_i^a to be closer to all images x_i^p than to any of the images x_i^n, as shown in (2).

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2, \quad \forall \left(f(x_i^a), f(x_i^p), f(x_i^n)\right) \in \mathcal{T} \tag{2}$$

Where $f(x_i)$ is the embedding of an image x_i, α is a margin that is enforced between positive and negative pairs, and T is the set of all possible triplets. Then, the network cost function to be minimized is described in (3).

$$\mathcal{L} = \sum_i^n \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \tag{3}$$

The overall architecture of the Triplet Neural Network is shown in Fig. 4.

B. *Reuse: Suggesting Tags Based on the Experience*

In the reuse phase, the best solutions are suggested from the cases retrieved in the previous stage. The most common method for this purpose is the weighted vote of the solutions proposed by the cases using their distance to the new case. In the case of a multi-label system such as the proposed one, one can either retrieve the k most voted solutions or use multi-label implementations of the Nearest Neighbors algorithm. Our proposal is to use the most voted solution, and then to search in the latent space for their k closest solutions using the Euclidean distance.

Given that the solution space is large, and that the problem to be solved such as recommending hashtags for a folksonomy is complex due to problems such as the user's freedom in defining the hashtags and the constant evolution of these hashtags, an alternative is

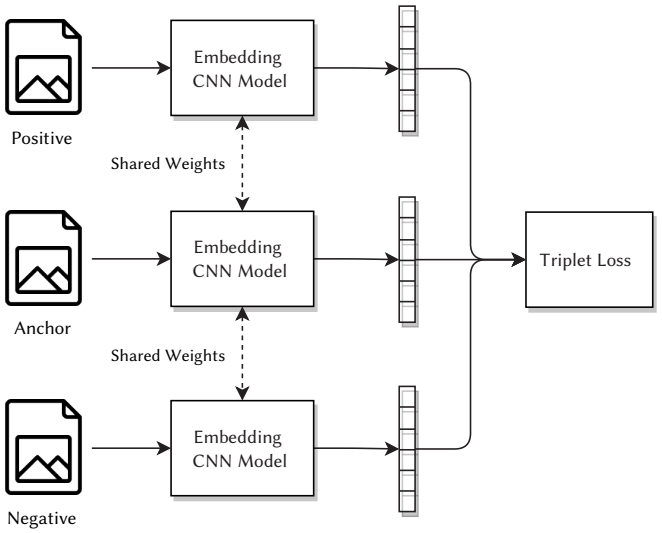


Fig. 4. Overview of the Siamese Neural Network architecture using triplet loss function.

proposed. Our approach consists of reducing the solution space to a set of semantically grouped clusters. In this way, hashtags belonging to the clusters obtained in the CBR reuse phase will be recommended.

1. Hashtags Latent Space

The hashtags recommended by our system are retrieved from a latent label space from the proposed solution. This latent space is formed by the embeddings of the hashtags built in the initialization of the system from the Case-Base. The label embeddings are obtained from a DNN. This DNN takes as input a tag and the ID of an image and identifies whether the tag and the ID are related or not. That is, given a tag t_i and an image i_i, $r(t_i, i_i) = 1$ if they are related, and $r(t_i, i_i) = 0$ otherwise.

The DNN architecture consists of two branches, one for hashtags and one for images. Each branch has an embedding layer of 64 components. This layer "encodes" the inputs to a feature vector, i.e. it uses the input as an index to obtain the corresponding feature vector. In each iteration, the weights of these vectors are adjusted obtaining at the end of the training a *NxM* matrix where *N* is the number of elements to encode and *M* the number of embedding components. Both branches are joined in a Dot layer that computes the dot product of the previous outputs. The next layers are a Reshape layer to resize the input to a one-dimensional vector and the last layer is a fully-connected layer. The hashtags latent space is formed by the weights of the hashtags embedding layer. An overview of this architecture is shown in Fig. 5.

Finally, the semantic clusters of the hashtags are obtained using the k-means algorithm. To determine the best number of clusters, the elbow method has been applied using the inertia (the sum of squared distances of samples to their closest cluster center) as metric.

*C. **Revise and Retain**: Adding the New Tagged Images to Case Memory*

In the last two phases of the CBR, the solutions are reviewed and the cases where the proposal was incorrect are stored. In a recommendation problem, in the review phase, the proposed tags are compared with those that the user actually used, so unlike most CBRs, no manual review is necessary. Those tags that the user finally used and the system did not recommend are stored in the retain phase. This allows our proposed system to do two things: on the one hand to learn from the tagging habits of the users, which in the case of folksonomies is changing; and on the other hand to add new tags to the system,

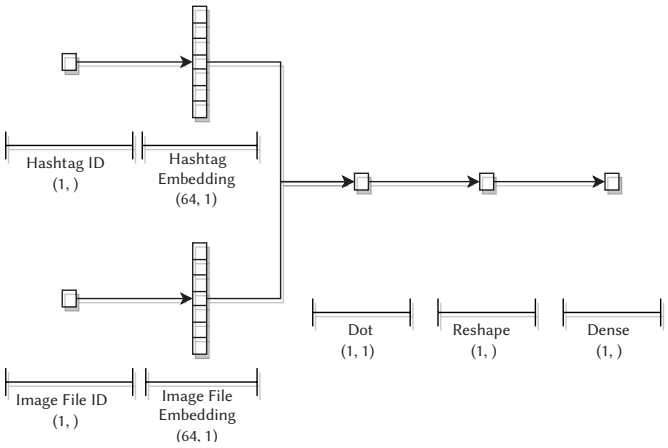


Fig. 5. Overview of the Deep Neural Network architecture for hashtag and image file embeddings.

which in the case of folksonomies the label space is very varied and divergent. Therefore, it is important that each time a new tag is added to the system, both the embeddings of the tags and their semantic clusters are recalculated.

V. Instagram as a Case Study

Since Instagram is currently the most relevant social network in the marketing field and that it is one of the most used and fastest growing social networks in recent years [1], this work validates our proposal with data obtained from a set of profiles of this social network.

Specifically, the dataset used in this case study arises from [36]. In this work, image and text are both used to categorize different Instagram influencer profiles according to their topics of interest. To this end, a dataset consisting of 10,180,500 posts from 33,935 Instagram influencer profiles is compiled. For each post, the dataset contains image files, captions, hashtags, usertags, number of likes, associated comments and other meta-data. Since it contains all the information necessary for the implementation of our proposal, this dataset is used in our work. In this case, the most relevant information of each post is related to the meta-data of the images and the associated hashtags. In besides applying a novel approach to this data, our goal is not to categorize user posts but to obtain and recommend a set of tags in order to optimize the dissemination of the image posted on Instagram.

In order to obtain the dataset on which to perform the tests, an undersampling of the original dataset was first performed. From the total number of posts we have randomly chosen a 5% sample, equally distributed for each of the topics. From that 5% we extracted the hashtags of each post, resulting in a total of 4,008,534 records. Many of the hashtags found in these records are hardly representative, so a filtering has been performed to eliminate those hashtags that have no more than 0.1% of representativeness. In this way, a final hashtag space of 2,083 hashtags was obtained, ranging from the most representative hashtag #ootd with 30,378 records to #ocblogger with 379 records. However, many hashtags are extensions of others, for example #recipevideo of #recipe. In order not to discard these extensions, they have been grouped under the root hashtag, i.e. for the above case #recipevideo has been replaced by #recipe. In Fig. 6 we can see the tag cloud with the representativeness of each hashtag. The final number of records in the dataset is 3,361,766. This dataset is divided into a training dataset and a test dataset in a ratio of 80/20.

As described in the previous section, the Case-Base of our CBR is constructed from the Case-Representations of the initial cases. The Case-Representations are embeddings of the images obtained from

Fig. 6. Influencers dataset hashtags wordcloud.

the CNN trained on the Siamese Neural Network. To train the Siamese Neural Network, a dataset of triples (anchor, positive, negative) has been constructed from the training dataset. The triplets have been created by taking the images as anchor and all those other images that use the same hashtags have been taken as positive, and those others that do not use any anchor hashtags have been taken as negative. That is, the relationship of the triplet (anchor, positive, negative) is whether or not the same hashtags are used. The Fig. 7 shows an example of image triplet.

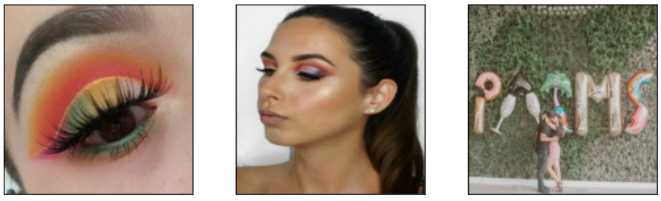


Fig. 7. Sample of image triplet CNN input. The first image is the anchor, the second image is the positive and the third image is the negative.

To obtain the latent space of hashtags, records (x_i, i_i, y_i) have been created where yi indicates whether the hashtag and the image are related ($y_i = 1$) or not ($y_i = 0$). This value is taken by the hashtag embedding DNN as a target. In the training dataset there are only positive relationships, so it has been necessary to create negative relationships, i.e. the relationship value is 0. For this purpose, all combinations of pairs pxi, iiq have been taken and those where there is no relationship have been taken as negative.

From this dataset, the latent space of hashtags was trained and obtained. An example of the hashtags closest to the hashtag #ootd in that space can be seen in Table I.

TABLE I. Top 10 Nearest Hashtags to Hashtag #OOTD From Hashtag Latent Space

Hashtag	Distance
#ootd	0.000000
#fashion	0.630845
#fashionblogger	1.063766
#style	1.126536
#outfitoftheday	1.863068
#outfit	1.884869
#styleblogger	2.210241
#instafashion	2.504243
#fashionista	2.607178
#instastyle	2.828929

Finally, the semantic clusters of the hashtags have been obtained from the latent space of hashtags. The algorithm used to obtain the clusters was k-means, using the elbow method to infer the best number of clusters. The results of the elbow method using inertia as a metric can be seen in Fig. 8 In the tests performed, for all possible values of k from 2 to 100, the best possible value was 14 clusters. In Fig. 9 a 2D projection of the latent space of hastags is shown, where all hashtags can be seen colored in red and colored in blue the hashtags closest to #ootd. The 2D projection of the latent space with the hashtags colored per cluster can be seen in Fig. 10.

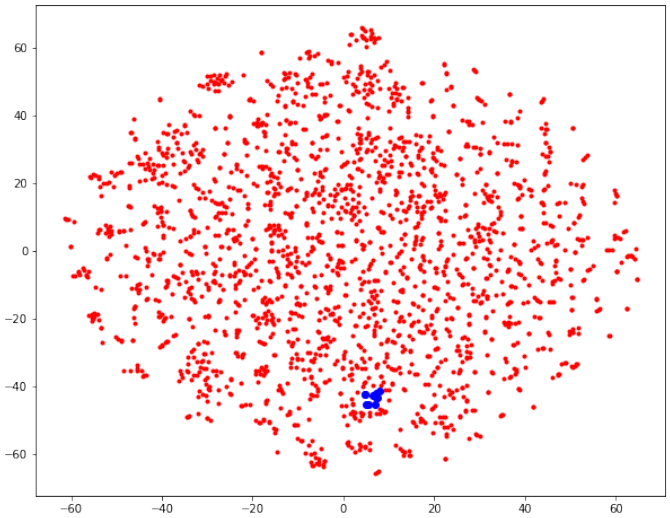


Fig. 9. 2D representation of the hashtag latent space using TSNE. Blue dots are the top 10 nearest hashtags to the hashtag #ootd.

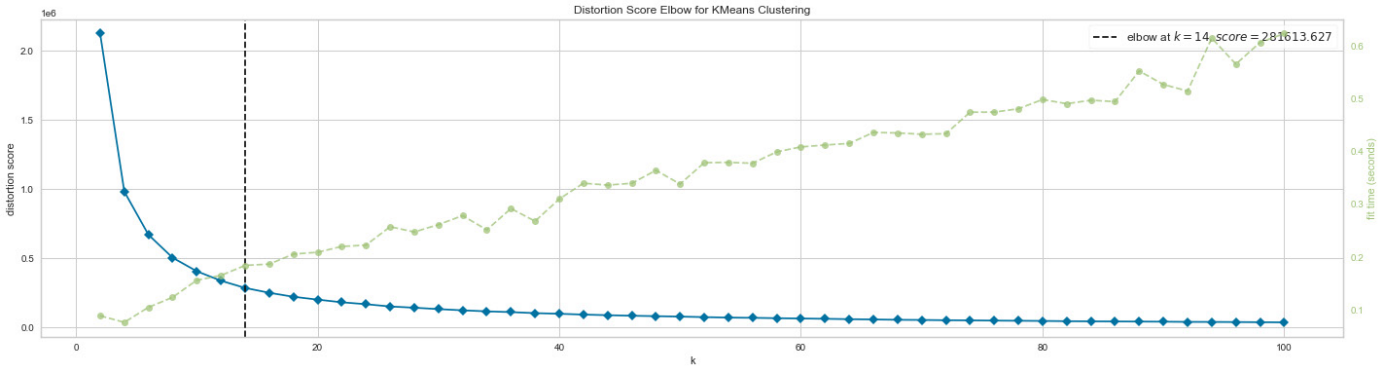


Fig. 8. Results of the elbow method for the number semantic clusters of hashtags.

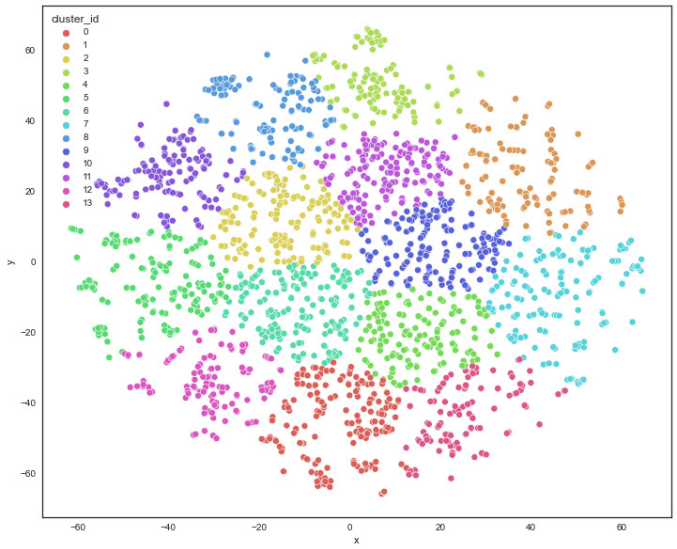


Fig. 10 2D representation of the semantic clusters of hashtags using TSNE.

VI. Results and Discussion

In this section we present the results of the tests performed. To compare our proposal we have chosen two other well-recognized methods also applied for feature extraction such as the color histogram and the Bag of Visual Words (BoVW). These two methods will generate Case-Representations against which to compare our proposal. All three methods use the same Case-Base filled from the Case-Representations of the training dataset. In the tests, a prediction of 10 hashtags has been performed for each new case. After several initial tests, the optimal value of nearest neighbors for the retrieve phase is 1000. In addition, since the number of cases in the Case-Base is initially very large and penalizes the query times in the retrieval phase, it has been reduced using Random Selection Undersampling method, reducing the cases taking into account the minority hashtag. This reduction of cases has not penalized the results and has improved the query times.

Unlike the metrics commonly used in recommender systems such as MAE and RMSE, we have chosen a different set of metrics as we find they are more suited to the purpose of the proposed system. The metrics chosen to evaluate the models were precision, recall and f1-score. Additionally, the distance between the gravitational center of the proposal's embeddings and the gravitational center of the user's hashtags' embeddings has been calculated. In this way, we can compare the quality of the model in terms of the semantic quality of the proposal.

In our tests we evaluate the overall matching of the proposed hashtags with the user's hashtags, the matching of at least one of the proposed hashtags with the user's hashtags, and the matching of the semantic clusters of the proposal with the semantic clusters of the user's hashtags.

TABLE II. Results for Full Hashtags Recommendation Match

Method	Precision	Recall	F1	Dst
Color hist.	0.0421	0.0602	0.0444	39.6797
BoVW	0.0479	0.0580	0.0439	38.8396
Proposal	0.0629	0.0696	0.0502	38.6142

The results of the first test can be seen in Table II. As we can see, our system improves the other two, both in precision and recall (and therefore in f1-score). This shows that our system not only makes fewer errors, but also hits more user hashtags. Moreover, the distance between the hashtags is smaller than in the other two, i.e., even if our system makes a mistake in the hashtag to recommend, it is semantically closer than the recommendations of the other two systems.

In the case of getting at least one of the recommended hashtags right, our proposal improves on the other two, as shown in Table III. As before, our model makes fewer errors and covers more right hashtags than the other two models. It is interesting to highlight that the color histogram in this case improves the BoVW, unlike the previous case. As a result, although BoVW has a higher precision and recall in the first case, it should do so in fewer recommendations than the color histogram. In other words, it has better metrics but does good recommendations in fewer cases.

TABLE III. Results for at Least One Hashtag Recommendation Match

Method	Precision	Recall	F1
Color hist.	0.1266	0.2044	0.1564
BoVW	0.1240	0.1451	0.1337
Proposal	0.1540	0.2216	0.1817

Finally, in the case where the user's recommendations and hashtags belong to the same semantic clusters, our system also outperforms the others (Table IV). In this case it reaches 22% precision and 33% recall, i.e., approximately one out of four recommendations matches semantically and the recommendations cover one third of the topics (within the folksonomy) that the user wants to tag in the image.

TABLE IV. Results of Hashtags Belonging to the Same Semantic Groups of Hashtags

Method	Precision	Recall	F1
Color hist.	0.2021	0.2776	0.2232
BoVW	0.2057	0.2733	0.2248
Proposal	0.2247	0.3269	0.2593

Although the precision and recall are low in general, we must keep in mind that we are evaluating a recommendation problem and that these values are usually low. We are not evaluating the exact prediction but how good the recommendations are. It should be taken into account that aspects such as the influence of the recommendation on the user when choosing hashtags cannot be evaluated a priori.

The results of the research experiments carried out in this article can be found at this link.

VII. Conclusions

This proposal has presented an intelligent system to suggest tags for an image previously submitted to social networks. Instagram tags are recommended based on the image features and previous experiences on other similar uploaded posts. Therefore, a CBR architecture that learns from previous solutions is applied. As a first step, the system is populated with tagged images submitted to the social network. Then, the system compares a new image manually selected with similar images stored in their memory. Finally, the recommendation of the system is a set of tags which helps to disseminate an image in a social network. Thus, this work addresses a multi-label problem.

In order to demonstrate the validity of the system and its independence of the dataset, a cross validation was carried out in order to evaluate the of the system of new tags.The overall results of the experiments carried out emphasize that the system can suggest concrete words as tags that influences in the visibility of the post. Another important point regarding the results is that, although the technique selected in our proposal to obtain the case representation is a neural network, the color histogram and the SIFT attributes grouped as BoVW could also be valid.

The use of folksonomies is a human-machine collaborative approach. Tags are automatically obtained from the image properties of those obtained for similar cases, but in addition, the tagging behavior of Instagram users is taken into account, so the algorithm adapts to new trends and randomness in the tagging process is reduced.

Words suggested as tags are always based on previous cases, so the system does not infer new knowledge based on the semantics of the words. For a future work, we propose the implementation of Natural Language Processing techniques in order to predict new words based on previous cases and the involvement of Instagram accounts to test the recommender in the social network.

References

[1] Statista, "Number of monthly active instagram users from january 2013 to december 2021," 2022. [Online]. Available: https:

[2] //www.statista.com/statistics/253577/ number-of-monthly-active-instagram-users/.

[3] S. Sood, S. Owsley, K. J. Hammond, L. Birnbaum, "Tagassist: Automatic tag suggestion for blog posts.," in *ICWSM*, 2007.

[4] M. Sohn, S. Jeong, J. Kim, H. J. Lee, "Augmented context-based recommendation service framework using knowledge over the linked open data cloud," *Pervasive and Mobile Computing*, vol. 24, pp. 166–178, 2015.

[5] W. Chen, Z. Li, "A study of tag-based recipe recommendations for users in different age groups," in *International Symposium on Emerging Technologies for Education*, 2016, pp. 315–325, Springer.

[6] D. Eck, P. Lamere, T. Bertin-Mahieux, S. Green, "Automatic generation of social tags for music recommendation," in *Advances in neural information processing systems*, 2008, pp. 385–392.

[7] O. Celma, "Music recommendation," in *Music recommendation and discovery*, Springer, 2010, pp. 43– 85.

[8] J.-H. Su, H.-H. Yeh, S. Y. Philip, V. S. Tseng, "Music recommendation using content and context information mining," *IEEE Intelligent Systems*, vol. 25, no. 1, 2010.

[9] A. Van den Oord, S. Dieleman, B. Schrauwen, "Deep content-based music recommendation," in *Advances in neural information processing systems*, 2013, pp. 2643– 2651.

[10] X. Wang, Y. Wang, "Improving content-based and hybrid music recommendation using deep learning,"in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 627–636, ACM.

[11] Y. Choi, J. Kim, E. Yun, S. Lee, D. Kim, "A new image search and retrieval system using text and visual features," in *WebNet World Conference on the WWW and Internet*, 2000, pp. 742–743, Association for the Advancement of Computing in Education (AACE).

[12] J.-W. Huang, C.-Y. Tseng, M.-C. Chen, M.-S. Chen, "Pisar: Progressive image search and recommendation system by auto-interpretation and user behavior," in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, 2011, pp. 1442–1447, IEEE.

[13] M. De Gemmis, P. Lops, G. Semeraro, P. Basile, "Integrating tags in a semantic content-based recommender," in *Proceedings of the 2008 ACM conference on Recommender systems*, 2008, pp. 163–170, ACM.

[14] P. Lopez-de Arenosa, B. Díaz-Agudo, J. A. Recio- García, "Cbr tagging of emotions from facial expressions," in *International Conference on Case-Based Reasoning*, 2014, pp. 245–259, Springer.

[15] S. Craw, B. Horsburgh, S. Massie, "Music recommendation: audio neighbourhoods to discover music in the long tail," in *International Conference on Case-Based Reasoning*, 2015, pp. 73–87, Springer.

[16] S. Nasiri, J. Zenkert, M. Fathi, "A medical case-based reasoning approach using image classification and text information for recommendation," in *International Work-Conference on Artificial Neural Networks*, 2015, pp. 43–55, Springer.

[17] M. B. Chawki, E. Nauer, N. Jay, J. Lieber, "Tetra: A case- based decision support system for assisting nuclear physicians with image interpretation," in *International Conference on Case-Based Reasoning*, 2017, pp. 108–122, Springer.

[18] D. López-Sánchez, J. M. Corchado, A. G. Arrieta, "A cbr system for image-based webpage classification: Case representation with convolutional neural networks," 2017.

[19] M. Wang, B. Ni, X.-S. Hua, T.-S. Chua, "Assistive tagging: A survey of multimedia tagging with human- computer joint exploration," *ACM Computing Surveys (CSUR)*, vol. 44, no. 4, pp. 1–24, 2012.

[20] J. Li, J. Z. Wang, "Real-time computerized annotation of pictures," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 6, pp. 985–1002, 2008.

[21] X. Li, C. G. Snoek, M. Worring, "Learning tag relevance by neighbor voting for social image retrieval," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 180–187.

[22] H. T. Nguyen, M. Wistuba, L. Schmidt-Thieme, "Personalized tag recommendation for images using deep transfer learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017, pp. 705–720, Springer.

[23] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, G. Stumme, "Tag recommendations in folksonomies," in *European conference on principles of data mining and knowledge discovery*, 2007, pp. 506–514, Springer.

[24] A. Hotho, R. Jäschke, C. Schmitz, G. Stumme, "Folkrank: A ranking algorithm for folksonomies," 2006.

[25] S. Lindstaedt, R. Mörzinger, R. Sorschag, V. Pammer, G. Thallinger, "Automatic image annotation using visual content and folksonomies," *Multimedia Tools and Applications*, vol. 42, no. 1, pp. 97–113, 2009.

[26] S. Lee, W. De Neve, K. N. Plataniotis, Y. M. Ro, "Map-based image tag recommendation using a visual folksonomy," *Pattern Recognition Letters*, vol. 31, no. 9, pp. 976–982, 2010.

[27] E. Amador-Domínguez, E. Serrano, D. Manrique, J. Bajo, "A case-based reasoning model powered by deep learning for radiology report recommendation," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 7, no. 2, 2021.

[28] M. Benamina, B. Atmani, S. Benbelkacem, "Diabetes diagnosis by case-based reasoning and fuzzy logic," *IJIMAI*, vol. 5, no. 3, pp. 72–80, 2018.

[29] O. R. Zaíane, "Building a recommender agent for e- learning systems," in *Computers in education, 2002. proceedings. international conference on*, 2002, pp. 55–59, IEEE.

[30] D. G. Lowe, "Distinctive image features from scale- invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[31] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, "Speeded- up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[32] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, "Orb: An efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE international conference on*, 2011, pp. 2564–2571, IEEE.

[33] J. Yang, Y.-G. Jiang, A. G. Hauptmann, C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, 2007, pp. 197–206, ACM.

[34] R. Chakravarti, X. Meng, "A study of color histogram based image retrieval," in *Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on*, 2009, pp. 1323–1328, IEEE.

[35] M. Hassan, C. Bhagvati, "Evaluation of image quality assessment metrics: Color quantization noise," *Evaluation*, vol. 9, no. 1, 2015.

[36] L. Torrey, J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI global, 2010, pp. 242–264.

[37] S. Kim, J.-Y. Jiang, M. Nakada, J. Han, W. Wang, "Multimodal post attentive profiling for influencer marketing," in *Proceedings of The Web Conference 2020*, 2020, pp. 2878–2884.

[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.K. Weiss, T. M. Khoshgoftaar, D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1– 40, 2016.

[39] E. Hoffer, N. Ailon, "Deep metric learning using triplet network," in *International workshop on similarity-based pattern recognition*, 2015, pp. 84–92, Springer.

[40] I. Melekhov, J. Kannala, E. Rahtu, "Siamese network features for image matching," *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 378–383, 2016.

Lucía Martín-Gómez

Lucía Martín Gómez has a PhD in Computer Engineering, an official master's degree in Intelligent Systems and a degree in Computer Engineering from the University of Salamanca. In the research context, she has made several scientific publications and has collaborated as organizing committee of some international conferences framed in multiple areas within artificial intelligence. She has participated in several research projects related to IoT, social network analysis, big data and Industry 4.0 at national and European level. In 2018 she was granted a Pre-doctoral Scholarship for research workers training awarded by the Junta de Castilla y León (España). She has developed her professional career working as a data scientist in big data projects, and is currently a Professor at the Pontifical University of Salamanca.

Javier Pérez-Marcos

Javier Pérez has a degree in Computer Engineering at the University of Salamanca, a master's degree in Intelligent Systems at the University of Salamanca, and is currently a PhD candidate at the University of Salamanca. He worked for two years as Researcher at BISITE research group at the University of Salamanca, and for three years as Data Scientist and Big Data Manager at Smart Internet Internacional Sl. He is currently working as Data Engineer at Frogtek Sl. In addition, he has been Visiting Professor at the University of Salamanca, Visiting Professor at the University of Deusto and for the last three years Associate Professor at the Pontifical University of Salamanca.

Rebeca Cordero Gutiérrez

Rebeca Cordero Gutiérrez holds a PhD in Logic and Philosophy of Science with a specialization in Social Studies of Science and Technology from the University of Salamanca. Her doctoral thesis was developed in the field of horizontal social networks and their business and social impact. She has been a university lecturer for more than 10 years and has participated as research staff in several competitive projects. Her lines of research focus on the impact of ICT and social networks in business and education. She has taught courses and seminars related to new technologies and business management. She is the author of several articles in relevant journals, and has participated in numerous national and international conferences.

Daniel Hernández de la Iglesia

Daniel Hernández de la Iglesia is a Ph.D in Computer Engineering, a Technical Engineer in Computer Systems and a Degree in Computer Engineering from the University of Salamanca. He has an official master's degree in Intelligent Systems and in recent years has been linked to different research groups in the field of artificial intelligence where he has participated in dozens of national and international research projects. He is the author of different book chapters and has presented more than twenty research papers in different international congresses. In addition, it has numerous scientific publications in international impact journals indexed in the JCR reference ranking. He has been awarded the first prize of the Open Data Contest organized by the Junta de Castilla y León (Spain) in 2013, and with the first innovation award for the best research project awarded by the Junta de Castilla y León (Spain) in 2016. He is currently a Professor at the Pontifical University of Salamanca.

An Event Mesh for Event Driven IoT Applications

Roberto Berjón*, Montserrat Mateos, M. Encarnación Beato, Ana Fermoso García

Universidad Pontificia de Salamanca, Salamanca (Spain)

Received 21 April 2022 | Accepted 1 July 2022 | Early Access 19 September 2022

Abstract

In IoT contexts, software solutions are required to have components located in different environments: mobile, edge, fog or cloud. To design this type of application, event driven architecture (EDA) is used to develop distributed, scalable, decoupled, desynchronized and real-time components. The interconnection between the different components is done through event brokers that allow communication based on messages (events). Although the design of the components is independent of the environment in which they are deployed, this environment can determine the infrastructure to be used, for example the event brokers, so it is common to have to make modifications to the applications to adapt them to these environments, which complicates their design and maintenance. It is therefore necessary to have an event mesh that allows the connection between event brokers to simplify the development of applications. This paper presents the SCIFI-II system, an event mesh that allows the distribution of events between event brokers. Its use will allow the design of components decoupling them from the event brokers, which will facilitate their deployment in any environment.

Keywords

Cloud Computing, CloudEvents, Edge and Fog Environments, Event Driven Architecture, Event Mesh, Internet of Things.

DOI: 10.9781/ijimai.2022.09.003

I. Introduction

CURRENTLY, all high performance IoT applications, regardless of their computing paradigm: Cloud, Edge, Fog [1], Edge mesh [2] and Cooperative-based systems [3], are developed from an event-driven architecture based on microservices. An IoT application based on microservices is structured through a collection of loosely coupled distributed components that facilitate the scalability and performance of the system. Moreover, the use of event driven architectures allows the real-time processing not only of a data stream coming from different data sources external to the application [4], but also of the data flow exchanged between the different microservices of the application.

There are two key elements to consider in these systems: how to represent the data to be processed and the communication channels through which to transport this data.

The data to be processed is represented in the form of an event. Through it, a series of other elements can be included that can be of great importance during its processing: its source, correlation, transactionality, etc. For this reason, the Cloud Native Computing Foundation (CNCF) promoted the CloudEvents specification [6] for describing event data regardless of the format (json, avro, protocol buffers, xaml) and protocol used for its transport (MQTT, AMQP, Kafka, HTTP, ...), thus guaranteeing its portability and interoperability. A CloudEvent contains two parts: data and metadata.

CloudEvent event data is the data represented through the event. It can be text or binary information. If it is text, the value is included in the "data" attribute of the cloud event. Conversely, if the data is binary, its Base64 encoded value is included in the "data_base64" attribute.

* Corresponding author.

E-mail address: rberjonga@upsa.es

CloudEvent event metadata provides contextual information. It is a set of mandatory and optional attributes in the form of a key value. Table I describes the attributes included in the specification. In addition, if necessary, applications can add new attributes.

TABLE I. CloudEvent Metadata Attributes

Attribute	Description	Category
source	Represents the identifier of the publisher app that broadcast the event. It is expressed as a URI.	Required
id	Event identifier. The sender of the event must ensure that two events from the same source must necessarily have different values for this attribute.	Required
type	A publisher broadcasts different types of events. This attribute (which is a string) indicates the type of event.	Required
subject	Identifies the context in which the source emits the event. Usually, a consumer subscribes to events broadcast by a given source and subject.	Optional
datacontenttype	Represents the content type of data value. It is a string in RFC 2046 format.	Optional
dataschema	It is a URI that identifies the schema in which the data is structured.	Optional
time	Represents the datetime at which the publisher broadcasted the event. RFC 3339 encoded string.	Optional

When integrating distributed systems, it is necessary to use event brokers through which data flows. Therefore, the components sending and receiving data are coupled with respect to the event broker used. As discussed above, one of the premises to be ensured when designing

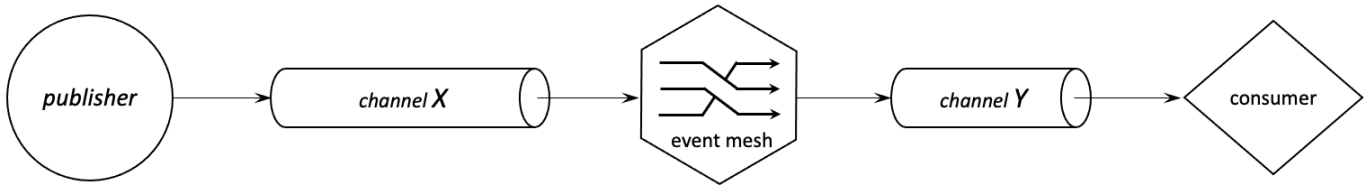


Fig. 1. Event mesh concept.

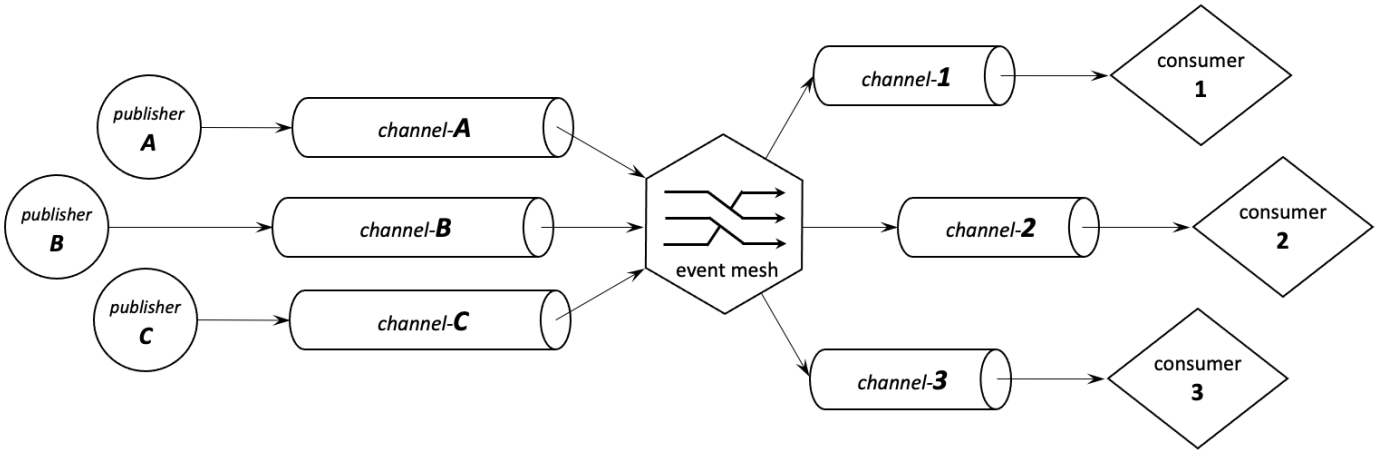


Fig. 2. Actual event mesh.

a distributed system based on microservices is the loosely coupling between its components. To ensure this, a middleware should be added to make the microservices independent of the communication channel through which data is sent and received. The simplest schematic of this type of middleware is shown in Fig. 1. It allows the publisher microservice to send data on the channel-X and the consumer microservice to receive data on another channel-Y. This intermediate layer connecting the event brokers is called event mesh.

An event mesh facilitates the simultaneous connection of different types of event brokers. It therefore acts as an event router. Depending on a set of rules, any event coming from any incoming channel can be forwarded through any of its outgoing channels. This allows great flexibility when integrating applications or extending the functionalities (adding new microservices) of an application. As can be seen in Fig. 2, consumer-1 is designed to receive events from channel-1. Through the event-mesh, channel-1 could receive events coming from channel-A, channel-B and channel-C issued respectively by publisher-A, publisher-B and publisher-C. All this without the need to modify the code of these components.

This is what SCIFI-II focuses on: the design and implementation of an event mesh that is able to connect different types of event brokers. Its mission will be to route incoming events to any of its outgoing channels. The events will be described through the CloudEvent specification. The event mesh will be configured through a dynamic set of rules. These rules will determine the route(s) to be followed by the event (to which outgoing channel to forward the event) based on the event characteristics (mainly described by its metadata).

II. Event Mesh Related Work

The following is a study of the event meshes currently available on the market and their main characteristics. Through this study, the disadvantages of these systems will be analyzed, and SCIFI-II will be presented as an alternative.

Knative Eventing [5]. It is an event mesh solution based on CloudEvents specifications [6]. Events can be redirected to any consumer if it has the ability to recognize and receive events using HTTP. Triggers are used to perform the event subscription, triggers are written using yml files, in which we indicate what values the CloudEvents event attributes should have, as a template.

The solution has two main disadvantages, the first one is that registration can only be done from addressable consumers applications over HTTP and the second is that the way applications show their interest in events is limited to the events meeting some characteristics related to their context properties, but not to the data they might contain.

Solace PubSub+ [7]. this event mesh solution is also an event broker that allows defining queues and hierarchical topics as destinations. Thus, a consumer or producer application can receive and/or send events to queues and hierarchical topics. Its main feature is the ability to register other event brokers to act as consumers or producers of these destinations, for example, an event that is directed to a Solace PubSub+ queue can be received in a MQTT topic.

The main limitation is that it is not an event centric solution, so a consumer is not self-sufficient in determining the type of events it wishes to receive. It is only a solution that connects event brokers to each other using previously configured destinations.

Argo Events [8]. It is a CloudEvents compliant solution. It incorporates a wide variety of “Triggers” (consumers) and “EventSources” (publishers). For the forwarding of events, we define “Sensors” that oversee forwarding the events (triggering of the “triggers”) when certain “dependencies” are met, which are related through “Conditions”. The characteristics of the events are determined by the dependencies from their data and/or their context.

The main limitations of this solution are its complexity in specifying the redirection rules (defining dependencies, creating conditions to relate the dependencies, including conditions within sensors) and the fact that its event sources do not include communication with mobile devices, despite including a greater number of event sources than other solutions analyzed.

TABLE II. Comparative Overview of Event Mesh Platforms

Platform	Event centric	CloudEvent compliant	Redirection based on	Rules format	Channels
Knative	Yes	Yes	Consumer rules	yaml	Based on http
Solace PubSub+	No	No	Static routes	-	Many except specific to mobile devices
Argo Events	Yes	Yes	Consumer rules	yaml	Many except specific to mobile devices
Serverless.com event Gateway	No	No	Topics	-	Based on servless http
Azure Event Grid	Yes	Yes	Azure triggers	code	Platform provided
Amazon EventBridge	Yes	No	Rules based on source and type events	Json	Platform provided
Oracle Events Service	Yes	Yes	Consumer rules	Json	Platform provided

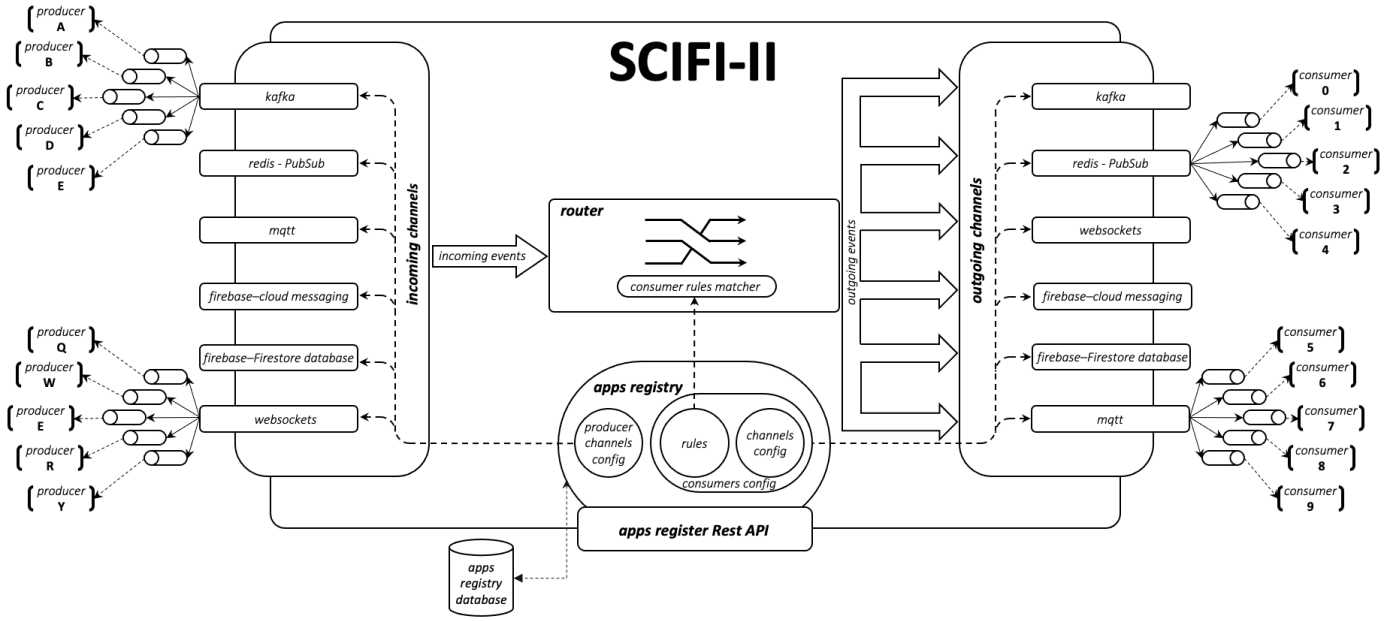


Fig. 3. SCIFI-II components.

Serverless.com event Gateway [9]. It is a serverless gateway that can redirect the events it receives (using HTTP request) to previously registered serverless applications. This solution allows to create sort of topics ("spaces") where event types are created. The way it works is as follows: when registering an application, it is necessary to indicate the event type and the space to which it subscribes, so that the events directed to these coordinates are routed to the registered serverless functions.

This solution is not an event mesh per se, although at the serverless level it can act as one. The main disadvantage found is that it is not an event centric solution since the subscription is made according to the destinations to which it is sent and not according to the characteristics of the events.

Other event mesh solutions provided with cloud platforms include the following: **Azure Event Grid** [10], **Amazon EventBridge** [11] or **Oracle Cloud Events Service** [12]. Their main limitation is that they are only cloud solutions designed to integrate the services of these platforms. Moreover, only Oracle Events Service and Azure Event Grid are CloudEvents compliant.

Table II summarizes the main characteristics of the analyzed platforms. As we can see, only the Argo Events platform is event centric, CloudEvent compliant, in which the redirection of events is done through rules specified by each consumer, and which supports multiple types of channels. Its main disadvantage is the excessive complexity with which these rules are described and the fact that it does not have channels for communication with mobile devices.

The following is a description of our SCIFI-II system that solves and simplifies the described limitations.

III. SCIFI-II

SCIFI-II is a reactive event mesh implemented using the Quarkus framework and therefore it is a cloud native system compatible with Microprofile Reactive Messaging standard specification. It allows dynamic registration of consumer and producer events applications. When registering a producer application, the event broker through which it emits the events must be indicated. As will be discussed later, SCIFI-II supports many types of event brokers. These events must be described using the CloudEvent specification. A consumer application must specify during its registration which events it wishes to receive and through which event broker. In order to specify the events, it wishes to receive, the consumer application must provide a set of rules. These rules must be enforced by incoming events so that they can be forwarded by the event mesh to the consumer. These rules are defined from the properties of the events, mainly from their metadata. SCIFI-II is therefore a CloudEvent compliant and event centric event mesh. A schematic of the different components of SCIFI-II is shown in the Fig. 3.

To dynamically register all applications, SCIFI-II includes a REST api. JSON is the format used to configure the application parameters. The application configuration data is stored in a Google Cloud Firestore database that feeds the apps registry module. This module is responsible for dynamically creating the source and sink connector instances linked to the incoming and outgoing channels of the applications. Additionally, it creates the rules that determine when an event must be redirected to a consumer application. All events received by the incoming channels are processed by the router module. This

module checks for each event the compliance with the rules defined by the applications. If a rule is fulfilled, router redirects the incoming event to the sink connector corresponding to the application that owns the rule.

As discussed above, in their registration, producer and consumer applications (or those playing both roles) must provide different data. Fig. 4 shows the json schema of the document to be provided in the registration of an application.

When a publisher is registered, the *incoming-config* property must indicate the channel on which it broadcasts its events. This attribute contains properties to define the necessary parameters for SCIFI-II to connect to that channel. When a consumer is registered, it is necessary to include two properties: *outgoing-config* and *rules*. Outgoing-config specifies the parameters required for SCIFI-II to connect to the channel to which the consumer is linked. On the other hand, the rules property defines the rules that the events must comply with in order to be routed to the consumer.

```
{
 "$schema": "http://json-schema.org/draft-07/schema",
 "$id": "http://scifi.upsa.es/register.schema.json",
 "type": "object",
 "definitions":   "config": {"type": "object",
                             : : : the object's structure depends of channel
                            }
                }
 "properties": { "id":   {"type": "string"},
                 "name": {"type": "string"},
                 "incoming-config": {"$ref": "#/definitions/config"},
                 "outgoing-config": {"$ref": "#/definitions/config"},
                 "rules": {"type": "array", "items": {"type": "string"}, "minItems": 1}
               },
 "anyOf": [ {"required": ["id", "name", "incoming-config"]},
            {"required": ["id", "name", "outgoing-config", "rules"]}
          ],
 "dependencies": { "outgoing-config": ["rules"],
                   "rules": ["outgoing-config"]
                 },
 "aditionalProperties": true
}
```

Fig. 4. Publisher or Consumer application register schema.

Fig. 5 shows the structure of the JSON documents needed to register both publisher and consumers. Of course, if an application plays both roles, the JSON would be a merge of both. SCIFI-II allows consumer applications to include additional properties which, as will be seen later, can be used in the definition of the rules.

Publisher

```
{
  "id": "id",
  "name": "name",
  "incoming-config": { channel definition }
}
```

Consumer

```
{
  "id": "id",
  "name": "name",
  "outgoing-config": { channel definition },
  "rules": ["rule-1", "rule-2", … "rule-n"]

  : : :  additional properties
}
```

Fig. 5. Publisher and consumer register.

The connectors of the different event-brokers supported by SCIFI-II are presented below.

A. Channels

This section explains the channels currently supported by SCIFI-II. They are classified according to the type of application that uses them. For each channel, the data to be provided in its registration is specified.

- ***sensors/actuators***: In IoT applications, MQTT is the main protocol used [13], in fact this protocol has become the lingua franca in the IoT world. The parameters that need to be included in incoming-config or outgoing-config for this channel are shown in Fig. 6.

```
{
  "type": "object",
  "properties": { "connector":          {"type": "string", "const": "scifi-mqtt"}
                  "mqtt.broker":        {"type": "string", "format": "uri"},
                  "mqtt.client-id":     {"type": "string"},
                  "mqtt.topic":         {"type": "string"},
                  "mqtt.qos":           {"type": "integer", "enum": [0, 1, 2]},
                  "mqtt.clean-session": {"type": "boolean"}
                },
  "required":["mqtt.broker","mqtt.client-id","mqtt.topic","mqtt.qos","mqtt.clean-session"]
}
```

Fig. 6. MQTT channel properties schema.

On the other hand, SCIFI-II also provides a connector for Kafka. Kafka is currently the leading distributed event streaming platform on the market. The configuration parameters are shown in Fig. 7.

```
{
 "type": "object",
 "properties": { "connector":         {"type": "string", "const": "scifi-kafka"}
                 "bootstrap-servers": {"type": "string"},
                 "application-id":    {"type": "string"},
                 "topics":            {"type": "string"},
                 "key":               {"type": "string",
                                       "enum": ["void",  "byte-array", "short",  "integer", "long",
                                                "float", "double",     "string", "uuid"]
               },
 "required":["connector", "bootstrap-servers", "application-id", "topics", "key"]
}
```

Fig. 7. Kafka channel properties schema.

- ***Mobile apps***: Mobile applications are an indispensable component in the development of IoT solutions for Smart cities [14] through the Mobile CrowSensing paradigm. In this sense, SCIFI-II provides connectors for different communication channels specific to mobile technology. One of them is Firebase Cloud Message. Through this channel it is possible to receive and send messages through the XMPP protocol to specific mobile devices (Fig. 8).

```
 "type": "object",
 "properties": {
   "connector":                  {"type": "string", "const": "scifi-firebase"}
   "firebase.type":              {"type": "string", "const": "messaging-xmpp"},
   "firebase.url":               {"type": "string", "format": "uri"},
   "firebase.project-id":        {"type": "string"),
   "firebase.api-key":           {"type": "string"),
   "firebase.credentials.file":  {"$ref": "#/definitions/credentials.file"),
   "firebase.recipientRules":    {"type": "array",
                                  "items": {"type": "object",
                                            "properties": {"test":      {"type": "string"),
                                                           "recipient": {"type": "string"}
                                                          },
                                            "required": ["test", "recipient"]
                                           },
                                  "minitems": 1
                                 },
   "required" : ["connector", "firebase.type", "firebase.credentials.file", "firebase.url",
                 "firebase.project-id", "firebase.api-key"]
```

Fig. 8. Firebase Cloud Message channel properties.

A message could also be sent to a topic. In this case all the mobile devices subscribed to it receive the message. For this use case, the channel could only be of type outgoing-config being their properties those shown in Fig. 9.

```
{
  "type": "object",
  "properties": {
                  "connector":        {"type": "string", "const": "scifi-firebase"}
                  "firebase.type":    {"type": "string", "const": "messaging-sdk"},
                  "firebase.url":     {"type": "string", "format": "uri"},
                  "firebase.credentials.file":   {"$ref": "#/definitions/credentials.file"},
                  "firebase.topic":   {"type": "string"}
                 },
  "required" : ["connector" , "firebase.type", "firebase.credentials.file", "firebase.url",
                "firebase.topic"]
}
```

Fig. 9. Firebase Cloud Message (topic target) outgoing channel properties schema.

Another of the connectors available to SCIFI-II in this area is to Google Cloud Firestore. Firestore is one of the main noSQL databases in the cloud that is used by all types of apps (android, Apple, or web apps). When an event is sent to this channel its event data contains in json format the new data to be added to a collection. The consumers of this channel receive events every time there is a change in the data of a collection (either when it is added, updated, or deleted). Its configuration properties are described in Fig. 10.

```
{
 "type": "object",
 "properties ": {
                  "connector":        {"type": "string", "const": "scifi-firebase"}
                  "firebase.type":    {"type": "string", "const": "database-firestore"},
                  "firebase.url":     {"type": "string", "format": "uri"},
                  "firebase.credentials.file":       {"$ref": "#/definitions/credentials.file"},
                  "firebase.firestore.collection":   {"type": "string"}
                 },
 "required":  ["connector", "firebase.type",  "firebase.credentials.file",  "firebase.url",
               "firebase.topic"]
}
```

Fig. 10. Google Cloud Firestore channel properties schema.

- ***Proxies***: In an application it is common to find components running before or after the data processor that perform data filtering and transformation tasks. Since one of the most used channels in these situations is Redis Pub/Sub, SCIFI-II also has a connector for it (Fig. 11).

```
{
  "type": "object",
  "properties": { "connector":     {"type": "string", "const": "scifi-redis"}
                  "redisauri":     {"type": "string", "format": "uri"},
                  "redis.channel": {"type": "string"},
                 },
  "required":  ["connector", "redisachannel", "redisauri"]
}
```

Fig. 11. Redis Pub/Sub channel properties schema.

- ***Web applications***: A typical use case in this context is the development of web applications to monitor and manage devices in real time. For this type of applications SCIFI-II provides a connector for WebSocket because it allows to establish a full-duplex asynchronous connection without the need for long polling (Fig. 12).

```
{
 "type": "object",
 "properties": {
                  "connector": {"type": "string", "const": "scifi-wsocket"}
                  "uri":       {"type": "string", "format": "uri"}
                },
 "required": ["connector", "uri"]
}
```

Fig. 12. WebSocket channel properties schema.

B. Rules

In SCIFI-II consumer applications must describe which events they wish to receive. To do so, they must indicate what features they should have. In the event these features are generally included in their context, which in the CloudEvents specification represents the event metadata.

During the registration of a consumer application, rules must be specified that incoming events must comply with in order to be forwarded to that consumer. These rules are defined by means of boolean expressions written with the Jakarta Expression Language (EL) syntax. For each incoming event, all the rules defined in the consumer applications are evaluated. The positive evaluation of any rule will result in the redirection of the event to the consumer application. In this way, an incoming event can be redirected to many consumer applications. Likewise, an incoming event may not be redirected to any consumer application.

In these ELs, the context bean "event" can be used to reference through it all the attributes included in the event metadata of the event. For example, the expression "event.type" shall refer to the value of the mandatory type attribute in the CloudEvent specification. For example, the rule in Fig. 13 will determine that the consumer application will be interested in events of type "es.upsa.scifi.dtwins.put" issued by "http://scifi.upsa.es/dtwins".

```
     (event.source eq 'http://scifi.upsa.es/dtwins')
and  (event.type   eq 'es.upsa.scifi.dtwins.put')
```

Fig. 13. Consumer rule with "event" bean context.

In the same way, these ELs can also reference the context bean "self". Through it, all the attributes included in the consumer application registry can be accessed. Remember that the json schema of this registry (see Fig. 4) allows additional attributes to be included as needed. For example, if the "from" attribute had been added to the consumer application record to represent the URI of the publisher app from which it expected to receive events, the rule in Fig. 13 could also be expressed as shown in Fig. 14.

```
{
  "id": "id",
  "name": "name",
  "outgoing-config": { channel definition },
  "rules": ["(event.source eq self.from) and (event.type eq 'es.upsa.scifi.dtwins.put')"],
  "from": "http://scifi.upsa.es/dtwins"
}
```

Fig. 14. Rule with "self" bean context.

The eventDataAsJsonObject() and eventDataAsJsonArray() functions can also be used in the rules. Through these functions it is possible to define rules based on the event data. The former is evaluated as the JSON Object contained in the event data of the event. The function checks that the CloudEvent event contains an event data and its datacontenttype is "application/json". Thus, through the value returned by this function, the properties contained in the event data can be accessed. The second function is like the first one, but in this case, it is evaluated as a json array, so each of its items can be accessed individually. Fig. 15 shows an example where a rule is created based on the "source" and "type" attributes of the CloudEvent context and the "temperate" attribute included in the json object representing the event payload.

```
     (event.source  eq self.from)
and  (event.type    eq 'es.upsa.scifi.dtwins.put')
and  (eventDataAsJsonOb ject() .temperature.doubleValue()   > 14.5)
```

Fig. 15. Consumer rule with eventDataJsonObject() function.

IV. Conclusion

This paper presents the SCIFI-II framework. This framework is an event mesh that allows the connection of multiple event brokers. Its use facilitates the development of distributed applications based on microservices and the integration of heterogeneous applications in a reliable and simple way. It adds a software layer that allows decoupling the event publisher and consumer components of a common event broker. In this way, each component can freely determine the infrastructure that suits it best.

Comparing this framework with others on the market, it is the most versatile as it is the only one with the following characteristics: event centric; CloudEvents compliant; event redirection is based on rules; the rules are represented through Expression Language, a powerful language that simplifies the creation of complex expressions based on the payload and context (or metadata) of the event; it also has a wide catalogue of channels from which to receive or send events, including those oriented towards mobile devices.

The SCIFI-II framework can be used for the development of applications based on event-driven architectures that can be deployed in both cloud and edge environments, as well as in legacy contexts.

References

[1] J. He, J. Wei, K. Chen, Z. Tang, Y. Zhou and Y. Zhang, "Multitier Fog Computing With Large-Scale IoT Data Analytics for Smart Cities," IEEE Internet of Things Journal, vol. 5, no. 2, pp. 677-686, 2018.

[2] Y. Sahni, J. Cao, S. Zhang and L. Yang, "Edge Mesh: A New Paradigm to Enable Distributed Intelligence in Internet of Things," IEEE Access, vol. 5, pp. 16441-16458, 2017.

[3] M. Rescati, M. De Matteis, M. Paganoni, D. Pau, R. Schettini and A. Baschirotto, "Event-driven cooperative-based Internet-of-Things (IoT) system," 2018 International Conference on IC Design Technology (ICICDT), pp. 193-196, 2018.

[4] P. Bellini, D. Nesi, P. Nesi and M. Soderi, "Federation of Smart City Services via APIs," 2020 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 356-361.

[5] Knative, "Knative Eventing,". Accessed: Mar. 06, 2022. [Online]. Available: https://knative.dev/docs/eventing/.

[6] "Cloudevents" Accessed Mar. 06 03 2022 [Online]. Available: https://

cloudevents.io.
[7] Solace, "Solace PubSub+". Accessed Mar. 06 03 2022 [Online]. Available: https://solace.com/.
[8] "Argo Events". Accessed Mar. 06 03 2022 [Online]. Available: https://argoproj.github.io/argo-events.
[9] serverless.com, "Serverless.com - Event Gateway". Accessed Mar. 06 03 2022 [Online]. Available: https://github.com/serverless/event-gateway.
[10] Microsoft Azure, "Azure Event Grid". Accessed Mar. 06 03 2022 [Online]. Available: https://azure.microsoft.com/en-in/services/event-grid.
[11] Amazon Web Services, "Amazon EventBridge". Accessed Mar. 06 03 2022 [Online]. Available: https://aws.amazon.com/eventbridge.
[12] Oracle Corp., "Oracle Cloud Events Service". Accessed Mar. 06 03 2022 [Online]. Available: https://www.oracle.com/cloud-native/events-service.
[13] B. Mishra and A. Kertesz, "The Use of MQTT in M2M and IoT Systems: A Survey," IEEE Access, vol. 8, pp. 201071-201086, 2020
[14] A. Capponi, C. Fiandrino, B. Kantarci, L. Foschini, D. Kliazovich and P. Bouvry, "A Survey on Mobile Crowdsensing Systems: Challenges, Solutions, and Opportunities," IEEE Communications Surveys Tutorials, vol. 21, no. 3, pp. 2419-2465, 2019.

Roberto Berjón

Roberto Berjón received his PhD. in Computer Science from the Universidad de Deusto in 2006. At present he is Professor at the Universidad Pontificia de Salamanca (Spain). He has been a member of the organizing and scientific committee of several international symposiums and has authored papers published in a number of recognized journals, workshops and symposiums. Nowadays he is member of the research group MARATON (Mobile Applications, inteRnet of things, dAta processing, semanTic technologies, OpeN data) where he currently focuses his work on IoT and mobile environments. At present time he is Program Director of the Master in Mobile Applications at the Universidad Pontificia de Salamanca.

Montserrat Mateos

PhD in Computer Science from Universidad de Salamanca in 2006. At present, she is a Professor at the Universidad Pontificia de Salamanca (Spain), and also, she is member of the research group MARATON (Mobile Applications, inteRnet of things, dAta processing, semanTic technologies, OpeN data) where she develops her research works in areas such as mobile technologies, IoT and Information Retrieval. She has been a member of the organizing and scientific committee of several international symposiums and has authored papers published in a number of recognized journals, workshops and symposiums. On other hand, she is external internship coordinator in Faculty of Computer Science.

M. Encarnación Beato

Mª Encarnación Beato (PhD.). Received a PhD. in Computer Science from the University of Valladolid in 2004. She is professor at the Universidad Pontificia de Salamanca (Spain) since 1997. At present she is a member of the MARATON (Mobile Applications, inteRnet of things, dAta processing, semanTic technologies, OpeN data) research group at the Universidad Pontificia de Salamanca. She has been a member of the organizing and scientific committee of several international symposiums and has co-authored papers published in a number of recognized journals, workshops and symposiums.

Ana Fermoso

PhD in Computer Science and Computer Engineering from the University of Deusto. She is currently Professor of Software Engineering at the Faculty of Computer Science of the Universidad Pontificia de Salamanca. She is a member of the MARATON research group, where she works on research lines of this group related to data retrieval, integration and processing, semantic technologies and open data, as well as mobile technologies and IoT. As a researcher, she is author and co-author of numerous scientific publications indexed in the main reference rankings (JCR and SCOPUS), she has participated as a presenter and has been a member of the scientific committee of numerous national and international scientific conferences. She has also participated in competitive research projects as principal investigator and collaborator and from which have derived intellectual property registrations of the software products developed in them. In addition, at present she is the Program Director of the Master's degree in IT project management and technological services. In regards to the topic of the master's degree, she has several certifications in the area. She is PMP (Professional Project Management) certified by the PMI (Project Management Institute), Scrum Master (PSM I) accredited by Scrum.org and by European Scrum, as well as ITIL4 Foundations certification for IT service management.

Board of Directors' Profile: A Case for Deep Learning as a Valid Methodology to Finance Research

César Vaca, Fernando Tejerina, Benjamín Sahelices *

University of Valladolid, Valladolid (Spain)

Received 22 April 2022 | Accepted 29 June 2022 | Early Access 19 September 2022

Abstract

This paper presents a Deep Learning (DL) model for natural language processing of unstructured CVs to generate a six-dimensional profile of the professional experience of the Spanish companies' board of directors. We show the complete process starting with open data extraction and cleaning, the generation of a labeled dataset for supervised learning, the development, training and validation of a DL model capable of accurately analyzing the dataset, and, finally, a data analysis work based on the automated generation of the professional profiles of more than 6,000 directors of Spanish listed companies between 2003 and 2020. An RNN-LSTM neural network has been trained in three phases starting from a random initial state, (1) learning of basic structures of the Spanish language, (2) fine tuning for scientific texts in the field of economics and finance, and (3) regression modeling to generate a six-dimensional profile based on a generalization of sentiment classification systems. The complete training has been carried out with very low computational requirements, having a total duration of 120 hours of processing in a low-end GPU. The results obtained in the validation of the DL model show great accuracy, obtaining a value for the standard deviation of the mean error between 0.015 and 0.033. As a result, we have been able to outline with a high degree of reliability the profile of the listed Spanish companies' board of directors. We found that the predominant profile is that of directors with experience in executive or consultancy positions, followed by the financial profile. The results achieved show the potential of DL in social science research, particularly in Finance.

Keywords

Artificial Intelligence, Board Diversity, Deep Learning, Finance, Long Short-Term Memory, Recurrent Neural Networks.

DOI: 10.9781/ijimai.2022.09.005

I. Introduction

THE amount of available open data from multiple areas has grown exponentially in recent years, which represents an enormous potential for researchers. However, due to the characteristics of much of the data generated, the processing of all this volume of information is also a huge challenge. Much of the data is characterized by its large size, high dimensionality and complex structure, making traditional econometric methods not entirely helpful, leaving a large amount of useful information behind. In this sense, the growing development of Machine Learning techniques provides researchers with effective tools for processing this type of data.

The generalization in the use of convolutional neural networks as well as multiple variants such as recurrent networks and long short-term memory, together with the use of normalization, regularization and optimization techniques, has allowed the creation of networks with many learning layers that can be trained effectively (References [1-4]). These deep neural network architectures (DL), together with the significant increase in available computing capacity thanks to the use of greatly improved GPUs, have allowed their successful application in new fields of knowledge in which they were previously not capable of generating positive results. Specifically, the progress achieved in the treatment of large series of data, of the type of text in the form of sentences and paragraphs, has given rise to sentiment classification applications that have shown great accuracy in different applications [4-8].

Our work represents a further step in this type of techniques, developing a model that generalizes sentiment classifier neural networks to achieve a regression of professional profiles in six dimensions and showing a specific application of it to the field of Economics and Finance, specifically, to the study of boards of directors. These governing structures have been the object of study in recent decades from different areas of knowledge, especially Finance, due to their important supervisory and advisory role in the company. However, the empirical studies carried out are mainly based on structured data, traditionally leaving aside a large amount of unstructured information due to the difficulties involved in its analysis.

Using data extraction and cleaning techniques, it has been possible to get unstructured data from the Annual Corporate Governance Reports of listed companies published on the website of the National Securities Market Commission (CNMV). These data are freely accessible to the public and describe the professional profile of the independent directors of the boards of directors of listed Spanish companies, i.e. those directors who are not contractually linked to the company nor are they shareholders of the company, so they are appointed on the basis of their professional background.

* Corresponding author.

E-mail addresses: cvaca@infor.uva.es (C. Vaca), ftejerina@efc.uva.es (F. Tejerina), benjamin.sahelices@uva.es (B. Sahelices).

The main contributions of the work are the following. In the first place, a methodology for extracting, cleaning and normalizing public data has been implemented that is repeatable and applicable to other conceptual environments. Secondly, starting from a state-of-the-art RNN-LSTM model for sentiment classification, it has been modified to convert it into a regression architecture so that a classification of professional profiles can be carried out in six dimensions. Thirdly, a manual labeling of a very significant set of directors has been carried out, which has allowed training the neural network and obtaining a very exact model with good generalization capacity. Fourth, a complete validation of the developed DL model has been carried out, showing its accuracy and validity for the study to be carried out. Finally, all this has allowed us to outline for the first time the professional profile of the boards of directors of Spanish listed companies. In general terms, the results show that the attribute most valued by companies when appointing independent directors is experience as an executive or consultant, followed by a purely financial profile.

The structure of the document is as follows: section II presents the state of the art of ML techniques and their possible applications to the field of Finance and Economics in general; section III describes the dataset and the model developed; section IV justifies and shows the results of the application of the model to the study of the profile of independent directors, while section V presents the conclusions.

II. Related Work

The Machine Learning (ML) model based on computational neural networks (NN), proposed many years ago, has experienced a great advance in the last 10 years due to the drastic improvement in processing speed by GPUs and new models, normalization, regularization, data augmentation and optimization techniques, among others. In the area of natural language processing, very significant advances have been made through the use of recurrent neural network (RNN) models that can handle arbitrarily large context sequences ([9], [10]). One of the most popular RNN models are long short-term memory networks (LSTMs), which have a better learning rate thanks to dropout, activation regularization, and temporary activation regularization [11].

Natural language learning is based on numerical vector representations of the text that allow capturing precise syntactic and semantic relationships. In [6], [7] the Skip-gram model is presented, which manages to establish syntactic and semantic relationships within the text with great efficiency and quality. In this way, it is possible to represent words and phrases, allowing generalization to represent and analyze documents efficiently. Document modeling is the base for a wide set of possible applications, starting with automatic text classification that allows extracting semantic information from text ([1]-[4], [12]). The most popular model used for the analysis and extraction of semantic information from a text are recurrent neural networks (RNN), and specifically the long short-term memory networks (LSTM) model, sometimes working together with convolutional networks (CNN) [11], [1]. There are multiple applications of text document modeling, including the classification based on feelings, the latent topic representation and the generation of information on the corporate culture of organizations ([13], [5], [8], [14]).

The word embedding model is used in conjunction with NN models such as RNN and CNN to perform high-level semantic analysis. In [14] corporate culture information is extracted in five dimensions, obtaining a correlation between its main corporate culture characteristics and business results and other significant corporate events. This study allows an analysis of a large number of corporations and the generation of accurate corporate information, which could not be done with traditional techniques. In the same vein, Hansen et al. (2018) [15] study how transparency affects the deliberation of monetary policy makers on the Federal Open Market Committee by analyzing its meeting records. The paper makes a methodological contribution by introducing latent Dirichlet allocation (LDA), a machine learning algorithm for textual analysis, to economic research. In asset pricing, Chen et al. (2014) [16] conduct textual analysis of articles and commentaries that were posted by investors on popular social media platforms and find that views expressed in both of them predict future stock returns. Likewise, Boudoukh et al. (2019) [17] analyze firm-specific news to predict volatility. Textual analysis is also actively used in Bellstam et al. (2021) [18], which uses the LDA method to examine analyst reports to specify a firm's level of innovation. A comprehensive review of the main applications of ML in finance can be found at Aziz et al. (2021) [19]. In general, major ML applications are in algorithmic trading, risk management and process automation.

III. Dataset and DL Model

The main objective of this work is the application of ML techniques to the analysis of the professional profile of the independent directors of companies listed in the Madrid Stock Exchange. CVs in the period 2003-2020 are used in free and unstructured format represented as simple sequences of text. The Deep Learning (DL) RNN architecture [9] is the most suitable for identifying dependencies in input text sequences. This type of network suffers from the vanishing gradient problem, which consists in the fact that the first levels of the network have a much lower learning rate than that of the last levels. The Long Short-Term Memory (LSTM) variant in the RNN architecture [1], [11] enables the propagation of the gradient to the first levels, thus improving its learning. The combination of the RNN architecture together with LSTM is perfectly adapted to the problem proposed.

The training of the DL model has been carried out in three phases. The first phase consists of creating a self-supervising neural network trained to recognize the structures of sentences and paragraphs written in Spanish, since this is the language used in the CVs. Using this network as a starting point, a second phase of training is applied to polish the understanding of sentences and paragraphs specifically in the financial, management, economic and business sectors. The objective of the second phase is to achieve a fine tuning of the neural network that allows for better results for the language and vocabulary structures used in the CVs.

The third training phase consists of using the encoder generated in phase 2 to carry out supervised training with a dataset labeled by a human expert to generate a regression model that characterizes each person using six profiles. The encoder contains the learning information of the grammatical structures of complete sentences in Spanish with an organization and vocabulary specific to the financial and business world. Using a supervised learning strategy based on manual labeling of a representative set of data, a regression neural network capable of linking free-form written language structures with six parameters that classify the manager in six profiles is generated.

A. Boards of Directors Dataset and Labeling

In order to ensure that the results of this work are consistent and repeatable, the decision was taken to train all the networks from scratch using datasets generated by us and to give up the transfer learning of networks trained with unknown data and parameters. For the first phase, we start from a network initialized with random parameters which we train to acquire knowledge of sentences and paragraphs in Spanish. The basic dataset used has been Wikipedia in Spanish [20] since the variability of its data, thematic diversity, writing styles and the breadth of vocabulary are guaranteed. Training with the total data set is not possible since the total volume of texts in Spanish

Wikipedia is very high. For this reason, a random selection of 10% of the total texts has been made, thus allowing training with viable computational requirements.

For the second phase of training, eight books have been selected from the thematic fields of finance, economics, business management, auditing, accounting and consulting. The goal is for the network to generate learning structures for sentences and paragraphs in these areas, as well as the corresponding vocabulary. The purpose of this dataset is to make a fine adjustment to our neural network, so that it can learn the differences between writing Wikipedia texts and scientific texts in the financial and managerial areas. It is expected to achieve a very efficient model in the prediction of words and phrases that constitutes a solid support for the third training phase [9].

Our proposal is related to the sentiment analysis and classification works [5], but we carry out a more detailed analysis using a regression model with six variables. The objective is to identify the profile of a director using the text of his CV provided in the company's public information. Based on sentiment classification systems, which have demonstrated their validity in different fields [5], their generalization is proposed to achieve a more detailed profile of directors. In phase three, six main activity profiles are identified: Financial (F), Executive/Consultant (E/C), Audit/Tax/Accountant (A/T/A), Legal (L), Political (P) and Academic (Ac). Based on these profiles, a regression-based model capable of assigning a numerical value to each of the profiles for each director is proposed.

The dataset consists of 137 firms listed on the Continuous Market of the Madrid Stock Exchange. Those belonging to the financial sector have been excluded, since their specific activity and regulations affect many characteristics of their governance, including the composition of their boards. The study covers the period 2003-2020. A total of 6561 director's profiles were analyzed. Most of the companies did not remain for the entire period but they were delisted at a given time or were successively incorporated.

The description of the directors' profiles (CVs) were taken from the Annual Corporate Governance Reports of the listed companies, which can be consulted on the website of the Comisión Nacional del Mercado de Valores (CNMV) in pdf format. Over the years, it can be seen that the description of the profiles has been expanding. In the early years, the descriptions were very brief or simply non-existent. Likewise, there are certain differences among companies in the degree of detail of such descriptions.

Of the 6561 director profiles, 1042 have been selected to obtain a training set on which manual labeling is performed, which is necessary for the supervised training stage of the regression model. The criterion for selecting the training set is a uniform distribution between the different types of profiles. Of these 1042 profiles, 100 have been separated, chosen randomly, to obtain a validation set that does not participate in the training and that allows us to perform a validation of the results. Additionally, the training set is internally divided into two groups to perform a standard training with a validation phase at the end of each epoch. This is used solely to measure the quality of the network's learning during its training and to make decisions about hyperparameters.

The human expert has labeled each director profile of the training dataset using the information in its public CV. For each of the six profiles, a number between 0 and 1 is assigned using a single decimal digit, that is, a standard assessment between 0 and 10 is made that estimates the weight of the corresponding profile within the CV. These profiles have not been considered mutually exclusive so maximum values (1.0) can be obtained in multiple and even in all profiles. For this reason, the criterion used in the labeling consists of assessing the CV data in a balanced way between the different directors and not making a percentage allocation. This allows a more objective and comparable assessment between the different directors. Given that the source of the data is a CV in the form of unstructured free text and in which the expression of similar merits can come from written expressions with very different structures and vocabulary, the labeling must be interpreted as an indicative fact of the professional profile and not as an exact numerical value.

B. DL Model and Training

The target of our DL proposal is to develop a statistical model of the language to estimate the distribution probabilities of the variations of linguistic units such as words, sentences, paragraphs and sets of paragraphs. This statistical model is built in two stages, the first for the training of the Spanish language and the second for the fine tuning of specific terminology. Taking this statistical model, we develop a second regression model that will allow us to carry out the multivariable profiling of the directors. To build the statistical model, a Recurrent Neural Network (RNN) with Long Shor-Term Memory (LSTM) [9] has been used. The main limitation of RNN models is that their learning capacity is small, which generates very long training times. To reduce this problem and increase the size of the analyzed context, the Long Short-Term Memory (LSTM) architecture is applied [1], [11]. This architecture uses memory units capable of storing information from multiple training sequences and, therefore, they are able to relate the information extracted from elements that are far apart from each other within the text sequence. This variant improves training times and greatly increases the length of the parsed context.

In the first phase, an RNN-LSTM network has been used to train the Wikipedia dataset and generate a base model with knowledge of linguistic structures in Spanish. In the second phase, this model has been refined using multiple books in the field of economics and finance. The encoder of this model has been used as a basic element for the training of a new RNN that has six regression outputs, each one to describe the profile of the directors, as explained in section 3.1.

As can be seen in the Figs. 1- 3 that describe the training in phases one (Wikipedia), two (Economics) and three (Regression), the network converges in a stabilized way. In phase one, no overfitting was observed, with the accuracy and perplexity values being 0.37 and 20, respectively, which shows a good relationship between the learning level and the training time. In phase two, overfitting appears, which was even more pronounced in previous training sessions and was reduced by applying a 30% dropout. Finally, in phase 3, a stable training with a reasonable value of RMSE was observed. Using a low-end GPU, the total computing time required to perform all three training phases was 120 hours.

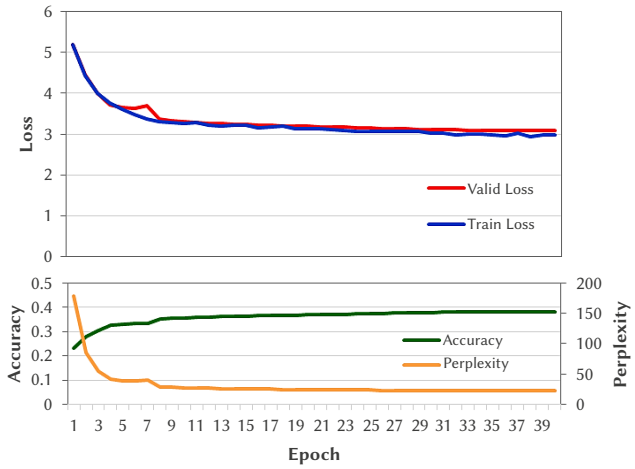


Fig. 1. Wikipedia Base Training.

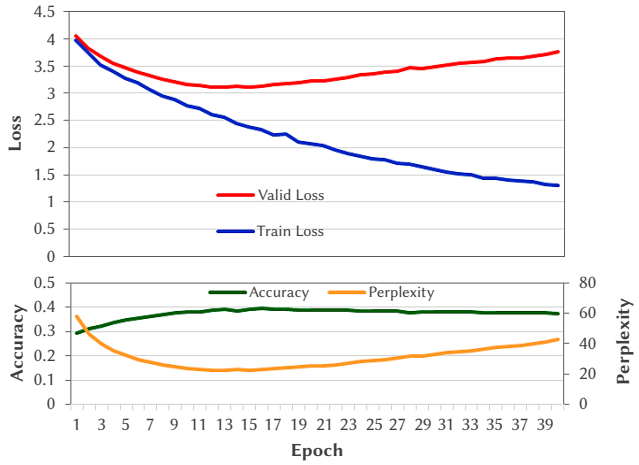


Fig. 2. Economy Base Training.

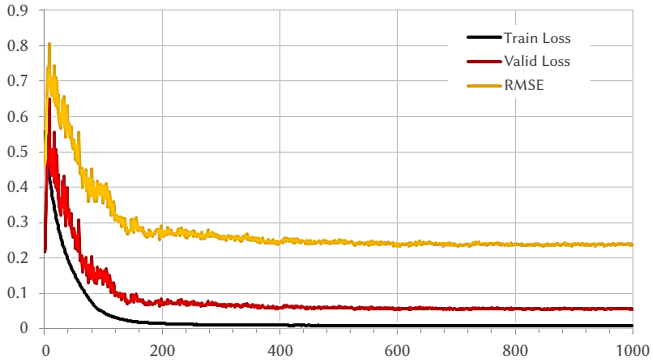


Fig. 3. Training for Regression.

C. Model Validation

To create a DL model that can be used to infer the total number of professional CVs available in the 2003-2020 period, it is necessary to carry out a prior validation phase that allows its behavior to be parameterized for correctly interpreting its results. In this sense, it is necessary to remember that the dataset used for the training, and therefore also in the inference, is an unstructured free text that describes the CV of the directors. Therefore, it is information of an imprecise nature which makes it difficult to generate results. The labeling of the dataset carried out by the human expert is also approximate and although numerical values are generated, these must be interpreted approximately as the quantification of the merits of a person referring to a specific profile.

We propose a validation process in two steps. In the first one, a numerical modeling of the model is carried out using labeled data which the network has not used for training. In this way it is possible to measure how this model mimics the behavior of the human expert. In the second step, a quantification scenario is designed to assign each of the six profiles to a category to measure the degree of accuracy with which our DL model identifies the profile of a director.

To carry out this validation stage, 100 professional profiles were randomly selected from the 1,042 that were labeled by the human expert. These 100 labeled CVs have not been used neither in training nor in the test carried out in each epoch. In other words, they have not participated in any way in the training of our DL model, nor have they affected decision-making about the hyperparameters. Therefore, this dataset allows us to carry out a clean and objective validation of our model.

The numerical behavior model of our neural network can be seen in Tables I and II which show the correlation analysis obtained by using the training and validation data respectively. As expected, our validation data has a higher error rate than the training data and provides us with a reliable indication of how well our network is able to generalize the trained knowledge. Our neural network model performs trend analysis in which the results are expressed as the average of large amounts of data corresponding to multiple profiles. As can be seen in Table II, the value of the standard error of the mean is small enough (between 0.015 and 0.033) to validate that the inference errors will in no case mask the results obtained. When performing the analysis with the 6561 profiles that make up our dataset, the value of the standard error of the mean is expected to be reduced to half of the values shown in Table II.

TABLE I. Correlation Analysis – Training Dataset

N = 942	*Professional Profile*					
	F	**E/C**	**A/T/A**	**L**	**P**	**Ac**
μ (error)	0.004	0.014	-0.008	-0.009	0.015	-0.003
σ (error)	0.097	0.129	0.083	0.099	0.046	0.092
corr. coef. (R)	0.971	0.962	0.957	0.961	0.986	0.962

TABLE II. Correlation Analysis – Validation Dataset

N = 100	*Professional Profile*					
	F	**E/C**	**A/T/A**	**L**	**P**	**Ac**
μ (error)	0.025	0.012	-0.037	-0.011	0.025	0.016
σ (error)	0.213	0.325	0.190	0.177	0.147	0.190
corr. coef. (R)	0.859	0.743	0.680	0.872	0.825	0.845
std. error of μ	0.021	0.033	0.019	0.018	0.015	0.019

Given the imprecise nature of the data used to train the network and given that the labeling performed by the human expert is, for the same reason, approximate, obtaining exact data from our model cannot be considered as the main target of our model. However, it is possible to assign a category to each of the six profiles to approximate their evaluation using a smaller number of intervals. Thus, it is possible to establish the profile using categories for each of the six values obtained in the regression. The discrete set used is made up of four categories that correspond to the intervals [0, 0.3), [0.3, 0.5), [0.5, 0.7) and [0.7,1.0] that are indexed by the integers 0, 1, 2 and 3. This reduction in the number of states is very useful for a profile analysis in which the data is imprecise, that is, it has a strong subjective component. Table III shows the success rate of the network for the training and validation datasets in which it can be seen that average success for validation data is 82.8%.

TABLE III. Quantized Categories Hit Rate

	Professional Profile					
	F	**E/C**	**A/T/A**	**L**	**P**	**Ac**
Training Set	93.8%	93.0%	96.9%	96.3%	97.1%	95.3%
Validation Set	77.0%	66.0%	92.0%	95.0%	83.0%	84.0%

Fig. 4 shows the confusion matrices for the validation dataset. It can be seen that the central categories (1 and 2) are much less represented than the limit categories (0 and 3), since in general the evaluation of a profile tends to values of the all/nothing type and the intermediate values are relatively infrequent. The results shown in this analysis validate our DL model for making scientific analysis of the profiling of boards of directors.

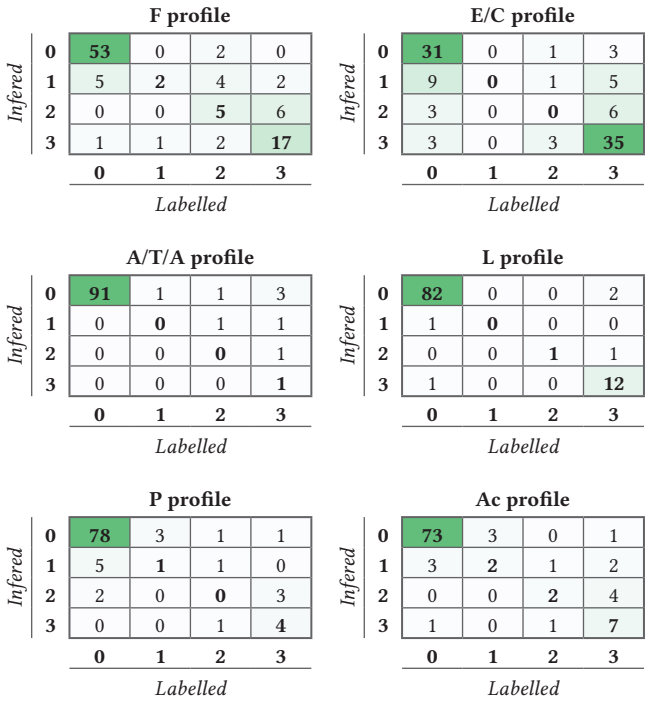


Fig. 4. Confusion Matrix for each profile (validation dataset).

IV. Profiling of Boards of Directors

A. Why Directors' Profile?

Sustainable finance refers to the process of taking environmental, social and governance (ESG) criteria into account when making investment or business decisions. In fact, corporate governance (represented by the G in the previous acronym) has been the subject of attention, both from the legislative perspective with the proliferation of codes of good governance and other types of soft-regulation, and from the academic field, for several decades. One of the features most discussed in this area in the financial literature is the independence of boards of directors, since it is considered essential to carry out their control function, as well as their influence on different company variables related to performance and value creation. In this sense, variables such as the percentage of independent directors, the size of the board, the separation between the figures of the chairman and the CEO or the number of meetings are analyzed (Hermalin and Weisbach, 2003 [21]). All of these would fall within the framework of what has been called structural diversity of the board.

However, in recent years boards are criticized for their excessive complacency and inability to prevent corporate crises, which has contributed to a widening of the perspectives from which they are analyzed. Thus, increasing attention has been paid in both academic and regulatory circles to board characteristics that can influence the effectiveness of the decision-making process. Such characteristics include, among others, the age, education, gender or nationality of the directors, grouped under the heading of demographic diversity of the board.

However, until now mainly those aspects of diversity that can be measured with quantitative and structured data (age, gender, years on the board, etc.) have been analyzed, leaving aside, due to the difficulty of collecting and structuring the information, aspects of great importance such as the professional profile of the board members. As Kim (2021) [22] points out, the analysis of unstructured data is one of the main fields of application of ML in finance research, although so far it is a largely unexplored field. It is only in recent years that attempts to apply ML to finance research have begun to emerge.

B. Directors' Professional Profiles

The analysis conducted allows us to understand the profile of the boards of directors according to the professional experience of each board member. As we pointed out in a previous section, we defined up to six professional profiles that, to a greater or lesser extent, have been considered in previous literature. They are as follows:

- Financial (F): Refers to those directors with experience in the financial sector, whether in banking institutions, any type of investment companies or the stock market in general. Güner et al. (2008) [23] and Booth and Deli (1999) [24] document that board members with financial expertise significantly influence firm financing and investment decisions.
- Executive/Consultant (E/C): Directors who have held or are currently holding different types of management positions in other companies or have carried out outstanding advisory tasks. These directors may have experience in different business sectors and management positions. According to the literature, independent directors help enhance firm value with their industry experiences (Drobetz et al., 2018 [25]; Faleye et al., 2018 [26]).
- Audit/Tax/Accountant (A/T/A): In this case, these are directors with specific expertise in auditing, tax or accounting. For instance, firms with accounting experts sitting on their audit committees show stronger accounting conservatism (Qiao, 2018 [15]).
- Legal (L): Lawyers and legal experts are classified in this profile. Their importance is highlighted by works such as that of Krishnan et al. (2011) [27], which shows that the presence of directors with legal backgrounds is associated with higher financial reporting quality.
- Political (P): Refers to directors who have held or are holding public offices of various kinds, especially political posts. The political profile can contribute to extending the company's relationships for its own benefit, although its presence has also been linked to sub-optimal decision-making. Houston et al. (2014) [28] find that the cost of bank loans is significantly lower for companies with board members with political ties, while Azofra and Santamaría (2004) [29] warned of the harmful effect of the presence of politicians on the efficiency of Spanish savings banks.
- Academic (Ac): Finally, this refers to those directors with academic experience. In this regard, some papers highlight that academic directors play an important governance role through their advisory and supervisory functions, leading to increased R&D performance and investment (Francis et al., 2015 [30]; Xie et al., 2021 [31]).

When appointing independent directors, companies will look for those profiles that best suit their needs. In general, independent directors perform the dual function of supervising and advising the board. The company will appoint them taking into account their ability to perform such functions on the basis of their professional profile. We assign a value between 0 and 1 to each profile for each of the independent directors analyzed. To the best of our knowledge, this is the first time that an overview of boards of directors based on the defined specific profiles has been presented.

C. Temporal Analysis

First, we examine the evolution of these profiles over time. The average number of independent directors per company fluctuates between 3 and 4, slightly exceeding this value in the last three years. Moreover, the number of companies in the sample tends to increase over the years. As a result, in absolute terms, an increase is observed in all the profiles considered, although from 2011 the "Political" and

"Legal" ones stagnated. The "Executive/Consultant" profile, on the other hand, is the one that has experienced the most marked growth in absolute terms.

Fig. 5 shows the relative importance of each profile over time. For each year, the sum of the different profiles is 1 (data have been normalized). The predominant profile in Spanish boards is by far the "Executive/Consultant" profile, with values starting at 0.35 and reaching 0.46 in 2020. The second most prevalent profile is "Financial", which is experiencing a slight but steady decline. The importance of the other four profiles is significantly lower.

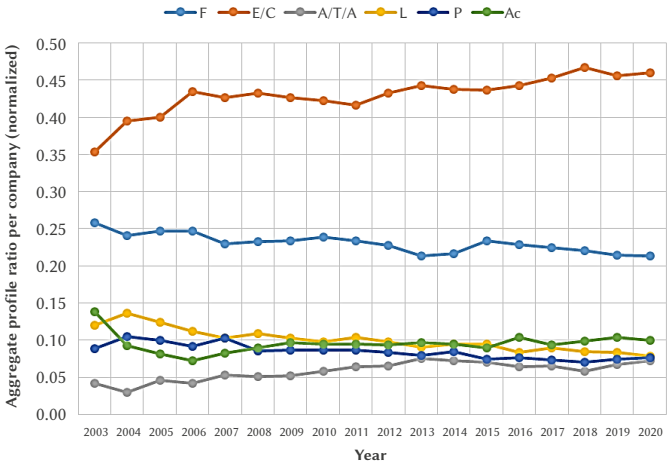


Fig.5 Aggregate Profile Ratios by Company per Year.

Regarding the evolution of each profile:

- "Executive/Consultant": after a pronounced increase, it experiences a slight decrease until 2011, from which a clearly upward trend is established.
- "Financial": the relative importance of this profile remains stable throughout the period, although with a downward trend. From an initial value of 0.26 to a final value of 0.21.
- "Legal" and "Political": of lesser relevance than the previous ones, they also experience a slight loss of importance throughout the period.
- "Academic": after a steep decline, it recovers from 2007 onwards with a stable trend.
- "Auditor/Tax/Accountant": this is the profile that has experienced the greatest increase over the years, although it is the profile with the lowest relative importance.

It should be noted that during the period in which the economic crisis was most pronounced (2007-2012), there was a certain stability in the "Financial" profile, with a slight upward trend. In turn, the "Audit/Tax/Accountant" profile experienced an increase in its relative importance.

D. Analysis by Sectors

To undertake this phase of the analysis we take the sectoral classification of the Madrid Stock Exchange, consisting of 7 sectors of activity (for the reasons given above, the financial sector, sector number 5, is not considered).

We reproduce the above analysis for each of the sectors considered (see Figs. 6-11). In all of them, the "E/C" profile remains the main one, followed by "F", although the relative importance of the latter varies for each sector.

TABLE IV. Sectors of the Madrid Stock Exchange

Sector 1	Oil and Energy
Sector 2	Basic Materials, Industry and Construction
Sector 3	Consumer Goods
Sector 4	Consumer Services
Sector 5	Financial Services
Sector 6	Technology and Telecommunications
Sector 7	Real Estate Services

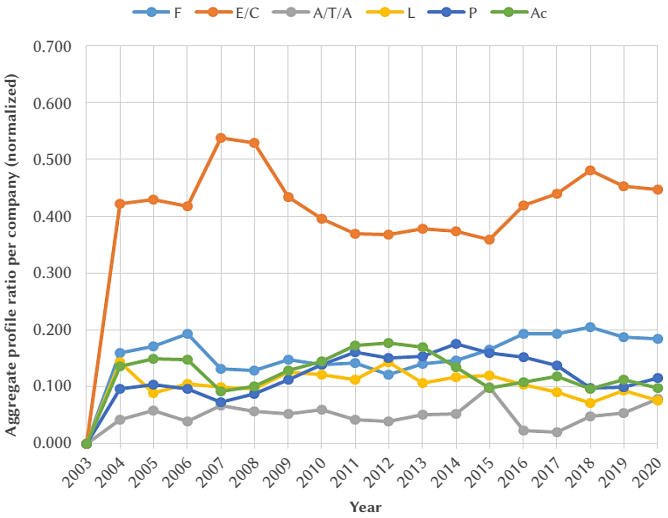


Fig. 6. Ratio of Aggregated Profiles by Company in Sector 1: The "E/C" profile is clearly the most important, although it declines sharply in the crisis period. It is noteworthy that the "F" profile is less important, especially during the crisis period, when it is overtaken by the "Ac" and "P" profiles, which grew considerably during this period. From 2012 onwards, the "F" profile became more important again, regaining second place in importance from 2015 onwards.

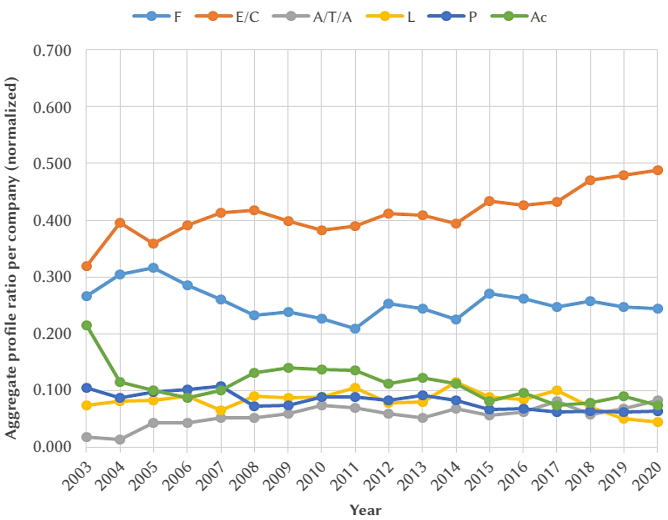


Fig. 7. Ratio of Aggregated Profiles by Company in Sector 2: Very similar to the general trend. There is a decrease in "F" during the crisis period and a slight recovery thereafter.

E. Discussion

The analysis carried out has allowed us to determine the profile of the independent directors with an acceptable degree of reliability. The preponderance of the "Executive/Consultant" profile seems to

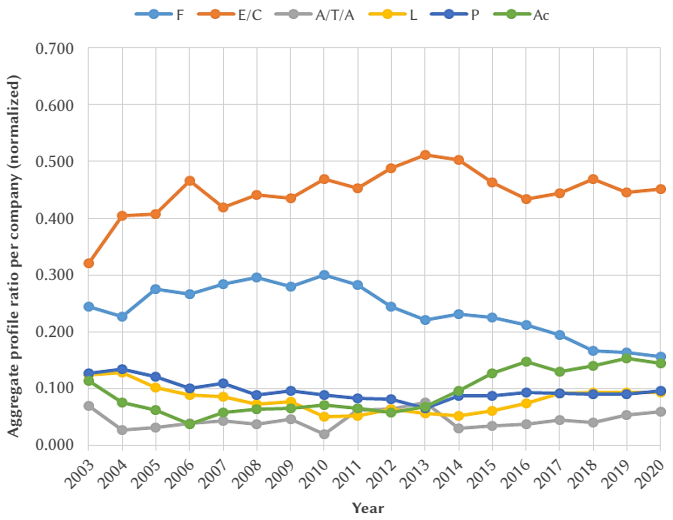


Fig. 8. Ratio of Aggregated Profiles by Company in Sector 3: Noteworthy is the sharp decline of "F" in the second part of the period in favor of the "Ac" profile, which from 2018 onwards practically equals the previous one.

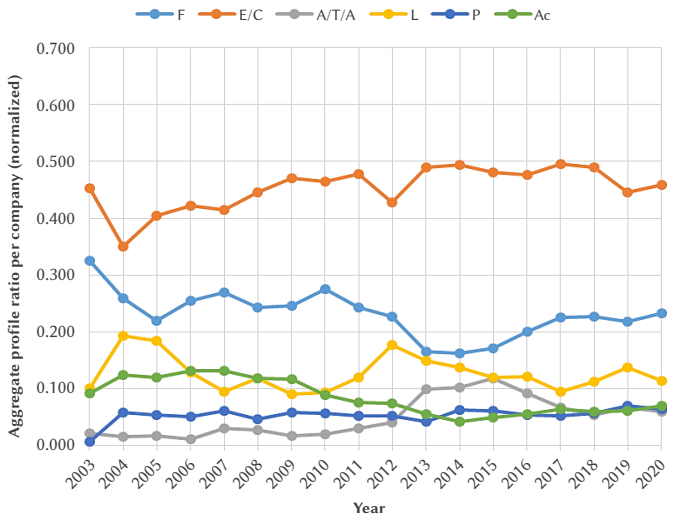


Fig. 9. Ratio of Aggregated Profiles by Company in Sector 4: The "E/C" profile continues to prevail. On the other hand, "F" has experienced a significant increase since 2015. In this sector we can highlight the relative importance of profile "L", whose relevance is significantly higher than for the total number of companies.

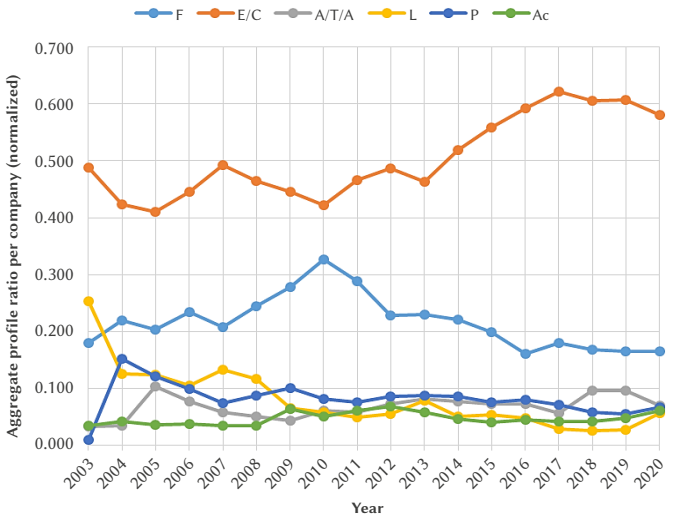


Fig. 10. Ratio of Aggregated Profiles by Company in Sector 6: After successive ups and downs, from 2010 onwards the "E/C" profile grows at the expense of "F". It is worth noting the initial importance of "L", which gradually decreases.

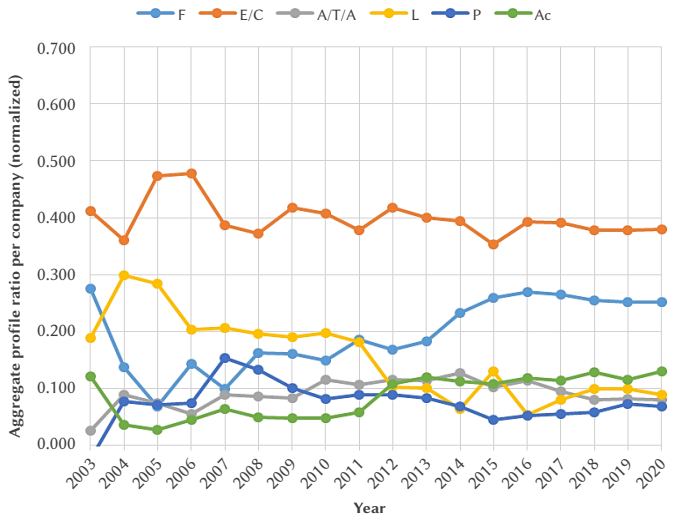


Fig. 11. Ratio of Aggregated Profiles by Company in Sector 7: From 2007 onwards, "F" increased, while "L" remained the second most important profile, ahead of "F" until 2011.

suggest that when appointing new directors, priority is given to their skills as advisors, giving less importance to their supervisory role. On the other hand, it would be expected that the "Financial" profile would increase during the years of greater economic crisis, given the financial problems it caused in a large number of companies. However, it can be seen that the relative importance of this profile remained stable and even showed a downward trend in the 2007-2012 period. Nevertheless, there was a strong increase in Sector 6 during the worst years of the crisis.

In this regard, the growing importance of the "Political" profile in Sector 1 during the crisis years is noteworthy. As this is a strategic and heavily regulated sector, the growing uncertainty caused by the crisis has encouraged the appointment of directors with this profile. Likewise, we highlight the role of "Academic" profile directors in Sector 3 since 2013. This is a profile that traditionally has not had much relevance, but which is recognized as being able to effectively perform the dual advisory and supervisory role required of boards of directors. Finally, the "Legal" profile stands out for its importance in different periods in sectors 4, 6 and 7, to some extent subject to legislative changes in the period under consideration. In this respect, it would be very interesting to look more deeply into the added value of these kinds of directors.

V. Conclusions

Advances in AI and, specifically, in DL allow for the development of learning models capable of analyzing information sequences of much greater length than previous methods were capable of successfully performing. This makes it possible to deal with complex problems in multiple areas of knowledge, and specifically, in the modeling of economic and financial data. Our work contributes to the progress of knowledge by presenting a model that in successive phases has been able to learn syntactic structures of natural language in Spanish from the area of economics and finance, and

use said knowledge to generate a regression model that accurately models professional profiles of boards of directors in six dimensions through a technique that can be considered a generalization of previous work on sentiment classification.

The results obtained present a more than reasonable degree of reliability, being an example of the capacity of ML techniques in the treatment of unstructured economic-financial information. In this field the application of this methodology is still at an incipient stage. However, given that a large part of the economic data is automated and growing exponentially, the application of these techniques will expand the frontiers of research in Economics and Finance.

Future lines of research are emerging from this work. In the area of extracting information from sequential data of the unstructured natural text type, it would be very interesting to delve into other hyperparameters and new models to increase the accuracy of the regression allowing to address new more complex problems. In the field of Finance, it would be useful to deepen and broaden the profiles considered in this work, as well as to analyze the influence of the different profiles on the firms' value creation process.

Acknowledgment

We would like to thank Manuel Astorgano Antón for his help in processing the data. This work was supported by the Spanish Ministry of Science and Innovation (AEI), Grants ID. PID2019-105660RB-C21 and PID2020-114797GB-I00. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] C. Zhou, C. Sun, Z. Liu and F.C.M. Lau, "A C-LSTM Neural Network for Text Classification," *The Journal of Machine. Learning Research*, vol. 12, pp. 2493-2537, 2015, doi: 10.48550/arXiv.1511.08630

[2] D. Yogatama, C. Dyer, W. Ling and P. Blunsom, "Generative and Discriminative Text Classification with Recurrent Neural Networks," 2017, arXiv 2017, 10.48550/arXiv.1703.01898

[3] A. Conneau, H. Schwenk, Y.L. Cun and L. Barrault, "Very deep convolutional networks for text classification," *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, Valencia, Spain, vol. 2, 2017, doi: 10.48550/arXiv.1606.01781

[4] S. Lyu and J. Liu, "Convolutional Recurrent Neural Networks for Text Classification," *Journal of Database Management*, vol. 32, no. 4, pp. 65-82, 2021, doi: 10.4018/JDM.2021100105

[5] Y. Kim, "Convolutional neural networks for sentence classification," *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, Doha, Qatar, 2014, doi: 10.3115/v1/d14-1181

[6] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, Tahoe, NV, United States, vol. 2, pp. 3111-3119, 2013, doi: 10.48550/arXiv.1310.4546

[7] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," *31st International Conference on Machine Learning, ICML 2014*, Beijing, China, vol. 4, 2014, 10.48550/arXiv.1405.4053

[8] B. Jiang, Z. Li, H. Chen and A.G. Cohn, "Latent Topic Text Representation Learning on Statistical Manifolds," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5643-5654, 2018, doi: https://doi.org/10.1109/TNNLS.2018.2808332

[9] W. De Mulder, S. Bethard and M. F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling," *Computer Speech and Language*, vol. 30, no. 1, pp. 61-98, 2015, doi: 10.1016/j.csl.2014.09.005

[10] M.G. Huddar, S.S. Sannakki and V.S. Rajpurohit. "Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN," *International Journal of Interactive Multimedia and Artificial Intelligence*. 2021;6(6):112-121. doi: 10.9781/ijimai.2020.07.004

[11] S. Merity, N.S. Keskar and R. Socher, "Regularizing and optimizing LSTM language models," *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, Vancouver, Canada, 2018, doi: 10.1109/TNNLS.2018.2808332

[12] P. Giudici, "Fintech Risk Management: A Research Challenge for Artificial Intelligence in Finance," *Frontiers in Artificial Intelligence*, vol. 1, 2018, doi: 10.3389/frai.2018.00001

[13] K. Li, F. Mai, R. Shen and X. Yan, "Measuring Corporate Culture Using Machine Learning," *The Review of Financial Studies*, vol. 34, no. 7, pp. 3265-3315, 2021, 10.1093/rfs/hhaa079

[14] S. Hansen, M. McMahon and A. Prat. "Transparency and Deliberation within the FOMC: A Computational Linguistics Approach," *The Quarterly Journal of Economics*, vol. 133, no. 2, pp. 801-870, 2018, doi: 10.1093/qje/qjx045.

[15] Z. Qiao, K.Y. Chen and S. Hung. "Professionals inside the board room: accounting expertise of directors and dividend policy," *Applied Economics*, vol. 50, no. 56, pp. 6100-6111, 2018, doi: 10.1080/00036846.2018.1489501.

[16] J. Boudoukh, R. Feldman, S. Kogan and M. Richardson. "Information trading and volatility: Evidence from firm-specific news," *Review of Financial Studies*, vol. 32, no. 3, pp. 992-1033, 2019, doi: 10.1093/rfs/hhy083.

[17] G. Bellstam, S. Bhagat and J.A. Cookson. "A text-based analysis of corporate innovation," *Management Science*, vol. 6, no. 7, pp. 4004-4031, 2021, doi: 10.1287/mnsc.2020.3682.

[18] S. Aziz, M. Dowling, H. Hammami and A. Piepenbrink. "Machine Learning in finance: A topic modelling approach," *European Financial Management*, pp. 1-27, 2021, doi: 10.1111/eufm.12326.

[19] https://es.wikipedia.org/wiki/Wikipedia

[20] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," *ACL. Association for Computational Linguistics* 2018, doi: 10.48550/arXiv.1801.06146

[21] B. E. Hermalin and M.S. Weisbach. "Boards of directors as an endogenously determined institution: A survey of the economic literature," *FRNBY Policy Review*, vol. 9, no. 1, pp. 7-26, 2003. doi: 10.2139/ssrn.233111

[22] H. Kim. "Machine Learning Applications in Finance Research," in *Fintech with Artificial Intelligence, Big Data, and Blockchain*, P.M.S. Choi and S.H. Huang Ed. Singapore: Springer, 2021, pp. 205-220.

[23] B. Güner, U. Malmendier and G. Tate. "Financial expertise of directors," *Journal of Financial Economics*, vol. 88, no. 2, pp. 323-354, 2008, doi: 10.1016/j.jfineco.2007.05.009.

[24] J.R. Booth and D.N. Deli. "On executives of financial institutions as outside directors," *Journal of Corporate Finance*, vol. 5, no.3, pp. 227-250, 1999, doi: 10.1016/S0929-1199(99)00004-8.

[25] W. Drobetz, F. Von Meyerinck, D. Oesch and M.M. Schmid. "Is director industry experience a corporate governance mechanism?," *SSRN Electronic Journal*, 2018, doi: http://dx.doi.org/10.2139/ssrn.2256477

[26] O. Faleye, R. Hoitash and U. Hoitash. "Industry expertise on corporate boards," *Review of Quantitative Finance and Accounting*, vol. 50, pp. 441-479, 2018, doi: 10.1007/s11156-017-0635-z.

[27] J. Krishnan, Y. Wen and W. Zhao. "Legal Expertise on Corporate Audit Committees and Financial Reporting Quality," *The Accounting Review*, vol. 86, no. 6, pp. 2099-2130, 2011, doi: 10.2308/accr-10135.

[28] J.F. Houston, L. Jiang, C. Lin and Y. Ma. "Political Connections and the Cost of Bank Loans," *Journal of Accounting Research*, vol. 52, no. 1, pp. 193-243, 2014, doi: 10.1111/1475-679X.12038

[29] V. Azofra and M. Santamaría. "El gobierno de las cajas de ahorro españolas," *Universia Business Review*, vol. 2, no. 2, pp. 48-59, 2004. Available at: https://www.redalyc.org/articulo.oa?id=43300204.

[30] B. Francis, I. Hasan and Q. Wu. "Professors in the Boardroom and Their Impact on Corporate Governance and Firm Performance," *Financial Management*, vol. 44, no. 3, pp. 547-591, 2015, doi: 10.1111/fima.12069.

[31] Y. Xie, J. Xu and R. Zhu. "Academic Directors and Corporate Innovation," 2021, Available at SSRN: https://ssrn.com/abstract=3954290.

César Vaca

César Vaca has a degree and a master's degree in physics from the Complutense University of Madrid. He has been a professor and researcher at the University of Valladolid since 1992. He is an expert in software development and complex data analysis. He is the author of the software environment that manages, organizes, optimizes and provides support for intelligent decision-making about all human medical resources in the Autonomous Community of Castilla y León in Spain. He has carried out multiple collaborations in the area of industry for the automation of highly complex systems, the creation of digital twins in the field of intelligent manufacturing and the application of Deep Learning in automatic quality control in industrial environments.

Fernando A. Tejerina

Fernando A. Tejerina Gaite, Ph.D., is an Associate Professor of Financial Economics at University of Valladolid (Spain). He has been Researcher Visiting Fellow at Leeds University Business School (UK), the University College Dublin (Ir), the Cass Business School (UK) and the University of Edinburgh (UK). His research is related to corporate governance, international business, family firms and innovation. He has earned research awards on five international organizations (University of Monterrey–Mx., JAAB-New York, Global Strategic Management Inc, European Journal of Arts and Sciences and Financial Studies Foundation). He has published several books and books chapters, and many papers in national and international journals. He has reviewed papers for many journals and conferences.

Benjamín Sahelices

Benjamín Sahelices has a PhD in computer science from the University of Valladolid, where he has been a professor and researcher since 1991. He has carried out stays and collaborations with research teams from the University of Tennessee in Knoxville, the University of Illinois at Urbana-Champaign and the University of Edinburgh, as well as in different Spanish universities. His main areas of interest are high performance computing, new memory hierarchies' designs, new paradigms of processing architectures, heterogeneous computing for deep learning and its applications. He has published in prestigious journals and conferences in the field of computer architecture. He regularly collaborates with research groups and companies in the area of industrial manufacturing, astronomy, economics, finance and biomedical engineering. He is a member of the GCME research group at the University of Valladolid.

Integrating Emotion Recognition Tools for Developing Emotionally Intelligent Agents

Samuel Marcos-Pablos[1]*, Fernando Lobato Alejano[2], Francisco José García-Peñalvo[1]

[1] Department of Computer Science, Universidad de Salamanca, (https://ror.org/02f40zc51), Salamanca (Spain)
[2] Faculty of Informatics, Pontifical University of Salamanca, Salamanca (Spain)

Received 21 April 2022 | Accepted 1 July 2022 | Early Access 19 September 2022

Abstract

Emotionally responsive agents that can simulate emotional intelligence increase the acceptance of users towards them, as the feeling of empathy reduces negative perceptual feedback. This has fostered research on emotional intelligence during last decades, and nowadays numerous cloud and local tools for automatic emotional recognition are available, even for inexperienced users. These tools however usually focus on the recognition of discrete emotions sensed from one communication channel, even though multimodal approaches have been shown to have advantages over unimodal approaches. Therefore, the objective of this paper is to show our approach for multimodal emotion recognition using Kalman filters for the fusion of available discrete emotion recognition tools. The proposed system has been modularly developed based on an evolutionary approach so to be integrated in our digital ecosystems, and new emotional recognition sources can be easily integrated. Obtained results show improvements over unimodal tools when recognizing naturally displayed emotions.

Keywords

Artificial Intelligence, Digital Ecosystems, eHealth, Emotionally Intelligent Agents, Human Computer Interaction.

DOI: 10.9781/ijimai.2022.09.004

I. Introduction

THE growing interest in the field of emotional recognition in the area of human-computer interaction (HCI) has fostered the development in recent years of numerous solutions aimed at making emotional recognition technologies available to inexperienced users. Although these technologies have been successfully employed in many fields such as online sales, trend analysis, or the study of user behavior in social networks, there is still a long way to go before they are fully developed and capable enough to be used in other fields.

The motivation of the present work is on the use of these emotional recognition tools for healthcare, and to incorporate them in our digital health ecosystems [1], [2]. There are many different applications for emotionally intelligent agents in healthcare. They can be used to help patients understand their emotional state, and to fill the information gaps when patients interact with health professionals helping health practitioners to increase the emotional understanding of their patients. Also, the emotional data gathered from the patient can be added to health records helping doctors in diagnosis and understanding of mental problems for example depression, schizophrenia, etc. Another application example are agents able to deliver personalized therapy. As more users employ technology to access health services, these services can be handled by intelligent agents. By endowing these agents with artificial empathy, we can increase the acceptance of users towards these new technologies.

* Corresponding author.
E-mail address: samuelmp@usal.es

However, the fact that many users still show reluctance to artificial intelligence (AI) means that special care must be taken when developing the agent's emotional intelligence, and even more in a delicate field as is human healthcare. The 'control degree' that emotional intelligence has over agent behaviors can make its actions better suited from an empathic interaction point of view but may generate unexpected behaviors leading to user rejection. However, there is a tendency towards the acceptance of AI which should increase as the users get more and more used to technology. Considering social robots as embodied agents, early studies in human-robot interaction in home environments suggested that users do not want a robot companion to be a friend, but to perform the tasks they are intended for. Contrary to these results, new studies suggest that robots able to accentuate their own personality are preferred by users [3].

Based on this tendency, the objective of this work is to integrate different commercially available emotion recognition tools into health services provision. However, current emotion recognition tools are generally unimodal, in the sense that they focus on a single communication channel (e.g., facial expressions, text, voice prosody, skin temperature, etc.) and provide an output which is associated to the activation of the universal emotions (namely: anger, fear, surprise, disgust, joy, and sadness). This approach is suitable for certain tasks that focus on specific affective aspects and not on the whole emotional spectrum (e.g., emotion recognition in an online review). On the other hand, multimodal emotion recognition by combining data collected from various communication channels can be employed for a wider range of applications, and provide surplus information with an increase in accuracy [4]-[7], therefore making it more suitable for health applications.

This paper presents an approach for multimodal emotion recognition using Kalman filters for the fusion of available discrete emotion recognition tools to be incorporated into emotionally intelligent agents. The proposed system has been modularly developed based on an evolutionary approach so to be integrated in our digital health ecosystems and allowing new emotional recognition sources to be easily integrated. The paper has been divided into five sections. The second section describes the background behind emotional intelligence and emotion parameterization for multimodal emotion fusion. The third section describes the proposed method and an implementation for testing purposes. The fourth section shows and discusses the results of the proposed approach. Finally, last section summarizes the main conclusions of this work.

II. Background

A. Emotionally Intelligent Agents' Main Characteristics

The different scientific disciplines where agents are applied means that different definitions of intelligent agents can be found in the literature. Originally robotics was the primary driver for agent-based research, however current agents include software mimicking or acting on behalf of humans (i.e., software agents) or internet robots (i.e., web-bots) [8]. From a general perspective, three major categories of agents can be distinguished: human agents, hardware agents and software agents [9]. Using the analogy of human agents, intelligent hardware and software agents are defined as capable of generating goals, performing actions, communicating messages, sensing environment, adapting to changing environments, and learning.

Continuing with the human analogy, emotional intelligence is defined as the accurate appraisal and expression of emotion in oneself and in others, the effective regulation of emotion in oneself and others, and the use of feelings to motivate, plan, and achieve day-to-day actions [10]. On the other hand, empathy can be defined as the cognitive ability to infer the thoughts and feelings of others and then developing a similar response to what the other person is feeling [11]. Empathy can be divided into cognitive empathy, or the ability to understand another's mental state; and affective empathy, described as the ability to respond with appropriate emotional reaction to the mental states of another.

From the above definitions, it can be inferred that endowing an agent with emotional intelligence involves sensing the environment, learning from the environment, and generating goals and actions adapted to the environment conditions and state. Therefore, emotional intelligent agents need to be provided with three main abilities: sense, compute, and act. In addition, as emotional intelligent agents are likely to interact with human users, their final goal should consider the environment with the focus on the user and provide a certain degree of empathy during human-computer interaction or collaboration. This means that the agent must be capable of capturing users' emotions (sense), appraisal of captured emotions to regulate its internal state (compute), and finally perform tasks where actions are regulated by the computed "emotional" state (act).

Many different approaches exist for capturing user's emotions, which include capturing movements [12], physiological parameters [13] or voice [14]. Deep learning is also widely employed for identifying face and body expressions in 2D and 3D [15]. Apart from trying to improve accuracy rates, recent systems focus on a multimodal approach to diminish the effects produced by variation of emotional display between users [15]. However, to date many of the non-invasive emotion sensing proposals (i.e., using cameras or microphones) encounter problems with emotion recognition in uncontrolled environments. For that reason, latest research is focused on capturing user's emotions under high variability conditions (light, noise, occlusions, etc.) [16].

In emotionally intelligent agents, as in human psychology, emotions are recognized as functional in decision-making by influencing motivation and action selection. Thus, emotion appraisal is needed after sensing to regulate the agent's internal state which in turn will determine the following actions to take. Emotions can be seen as a response to a certain stimulus that elicits a tendency towards a certain action, and as complex feedback signals that shape behavior. Therefore, processing emotions should be approached from a dual perspective: motivated action and feedback. Therefore, emotional intelligence in software agents can be seen as how the system processes emotions, focusing on how input is translated through an algorithm to an output and whether it contains a knowledge of past events or history. There are several types of algorithms for this purpose, which include fuzzy models, Markov models, neural networks, probability tables, reinforcement learning and unsupervised machine learning approaches such as K-means, K-medoids or self-organizing maps [3].

Acting is related to the tasks or operations the agent oversees conducting. For emotionally intelligent agents, actions should be adapted to the environment conditions and state. This means that not only agent's actions should be derived from the users' emotional state, but preferably the agent should return the user corresponding emotional feedback. The way such emotional reaction is expressed highly depends on the agent's degree of anthropomorphism, and can be as simple as a text message, color, or sound. However, as the agent anthropomorphism increases (e.g., in virtual avatars or robots), it turns necessary to employ natural communication channels (e.g., voice synthesis, facial animation) to match the agent's behavior with its appearance and avoid falling into the uncanny valley [3].

From the three characteristics described above, this paper focuses on capturing users' emotions (sense) using a multimodal approach based on different emotional sources.

B. Emotion Parametrization for Multimodal Recognition

Recognizing emotions from sensed signals is a complicated task but can be divided into two main steps: feature extraction and classification. During feature extraction, signal processing techniques are employed to produce a set of numerical values from the sensed signal (i.e., audio, video, EMG, and others). These numerical values collect certain features of the signal so that they can be processed by a computer. The extracted features are then processed by a classifier, which is complemented by a scoring function to produce the final emotional estimation. After feature extraction, emotion classification aims to provide with meaning to the observed features.

Multimodal emotion recognition can be seen as the fusion of information from different sources. There are two main approaches to fuse emotional data: feature-level fusion or early fusion, and decision-level fusion or late fusion [4]. Feature-level or early fusion fuses the features extracted from various sources such as visual, text and audio features into a unique feature vector which is sent for analysis. As the correlation between the features is performed at an early stage, feature level fusion can provide better results. However, putting all features in the same format is not an easy task and can induce to errors, as the features obtained from different channels can differ in many aspects. On the other hand, in decision-level or late fusion approaches the features of each input channel are examined and classified independently, and then their results are fused to obtain a decision vector as an output. The advantage is that the fusion of decisions obtained from different tools is easier as these tools usually have the same form of data. Additionally, each input channel can be processed with the most suitable classifier for its particular features.

Whichever approach is taken for computerized emotion recognition (unimodal, feature-level or decision-level fusion), it is necessary to parameterize human emotions, so that recognition can be done using quantitative computational techniques. In general terms, emotion parameterization can be divided into simple and dimensional approaches. Simple discrete models associate a set of detected patterns in the sensed emotional sources to the basic core of 6 universal emotions (namely: anger, fear, surprise, disgust, joy, and sadness). Such patterns do not need to be related with the psychology behind the expression, and unimodal recognition tools and many feature-level techniques use this approach.

On the other hand, dimensional approaches parameterize emotions as a lineal combination of different psychological dimensions. Emotion parametrization has been widely studied in psychology, and the most accepted dimensional models of emotion are the circumplex model, the vector model , and the Positive Activation - Negative Activation (PANA) model [17]. Multimodal decision-level emotion recognition systems employ this dimensional approach.

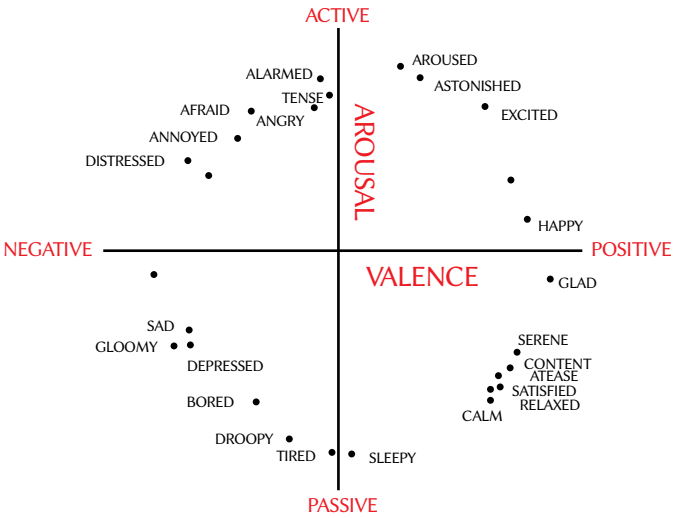


Fig. 1. The circumplex model of emotions.

In the circumplex model of affect, emotions can be categorized by 2 dimensions: valence, from unpleasant (negative) to pleasant (positive); and arousal, from passive (weak emotion) to active (strong emotion). By varying the values of each dimension, emotions can be plotted on two coordinate axes (see Fig. 1). Emotions are distributed in space with dimensions of arousal and valence in a circular pattern centered on medium arousal and neutral valence. In the vector model valence is modeled as binary, so emotions are plotted in a v-shape around the positive valence axis. The PANA model is like a 45-degree rotation of the circumplex model, where the two axes are: Positive Activation (PA), which goes from active, elated, and excited, to drowsy, dull, and sluggish; and Negative Activation (NA), which goes from distressed, fearful, nervous to calm, at rest, and relaxed. Vector models have been found to better describe emotional properties of text, whereas circumplex and PANA models have been identified for describing emotion in words, prosody and facial expressions [17].

The following section presents our approach for multimodal decision-level emotion recognition using Kalman filters for the fusion of decisions obtained from different commercially available recognition tools. As these decisions come from different communication channels, the circumplex model is used to parameterize emotions.

III. Methods

A. Acquiring and Parameterizing Existing Emotional Tools

In order to integrate different emotional recognition sources, we have developed a modular architecture which was introduced in [18]. The adopted approach was intended to be seamlessly deployed not only in computational agents, but also in physical agents such as robots. It consists of two main submodules: the facial emotion recognition submodule and the speech emotion recognition submodule, although additional emotion recognition sources can be added. Data coming from different sources is incrementally fused employing Kalman filters, as will be described in the next section. In addition, the developed architecture can be applied to physical agents such as robots, as data is exchanged between the different submodules employing ROS (Robot Operating System) messages.

The emotional recognition system modules have been programmed using both JavaScript and Python. Sound and image capture are performed through JavaScript, along with the different calls to cloud emotional APIs for image recognition and speech-to-emotion transcription. On the other hand, audio splitting and Kalman filtering have been deployed in Python. Python has also been employed to implement some of the message interchange over ROS. In addition, a front-end web page for testing purposes has been developed in html (Fig. 2).

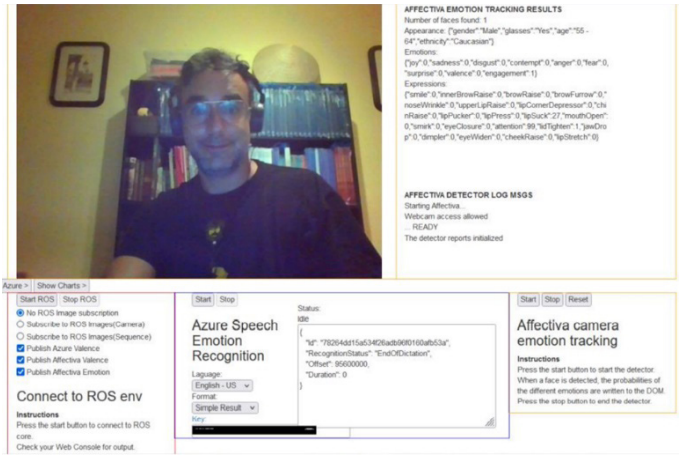


Fig. 2. Developed html interface for testing the proposed architecture.

At present, facial emotion recognition is approached by the combination of two tools: Affectiva's Affdex SDK and Microsoft's Emotional API. A call to these emotion recognition services takes a video or an image as an input and analyzes its emotional content returning a set of values using JSON syntax. After parsing the returned JSON, a value ranging from 0 to 100 is obtained for each of the six universal emotions (anger, disgust, fear, joy, sadness and surprise) [19] plus contempt and neutral. This value indicates the "activation" level for each expression.

Rather than using this value, and to integrate the facial output with other submodules, we have taken a circumplex approach to emotion recognition (See Fig. 1). Although the circumplex model is a continuous model where primary emotions can be expressed at different intensities and can mix with one another to form different emotions, there are studies in the literature that have tried to assess the approximate location of these emotions in the valence/arousal axis [20], [21]. Based on these works, we have converted the returned activation value of each expression into two components ranging from [-1, 1] for both valence and arousal:

$$A_E = 2\alpha_E(\frac{Val_E}{100} - 0.5) \quad (1)$$

$$V_E = 2\beta_E(\frac{Val_E}{100} - 0.5) \quad (2)$$

Where A_E and V_E are the valence and arousal of the expression 'E', whereas Val_E is the returned activation value from the facial recognition tools for that expression. The components (α_E, β_E) have been approximated from the literature as follows: anger (0.8, -0.8); disgust (-0.5, -0.5); fear (0.4, -0.9); joy (0.8, 0.3); sadness (-0.8, -0.2); surprise (0.95, 0.0). As the recognition focuses on the six universal expressions, we have discarded contempt and considered neutral as (0, 0).

A similar approach is followed for speech emotion recognition tools. Namely, current implementation makes use of Vokaturi prosodic-acoustic emotion recognition and Microsoft's Speech-to-Text and Text Analytics. Audio is captured continuously and processed using the SoX – Sound eXchange audio editing software. When a silence in speech is detected, the audio file chunk is sent to the prosodic emotion recognition module. The output of this module is then converted into valence and arousal following a similar approach as the one described before for the facial emotion recognition tools.

On the other hand, emotion recognition in the speech content analysis is divided in two main stages: speech to text, where the audio signal is converted into words, and text sentiment analysis, where text is provided with emotional meaning. After a chunk of raw audio data is captured, it is sent to the cloud service. The cloud service responds with a JSON containing the recognized text, along with other parameters such as the detected language, recognition confidence or the duration of the speech. The second step takes as an input the speech transcript, calls the Microsoft text sentiment analysis and returns a sentiment score which ranges from 0 (which represents a negative sentiment) to 100 (representing a positive sentiment). This sentiment score is translated into a valence value ranging from [-1, 1].

B. Integrating Emotional Sources

To integrate the data computed from the different emotional recognition sources, a sensor fusion approach is taken. Sensor fusion or data fusion is widely used in other fields such as robotics, where data coming from different sensors is merged to increase accuracy and reduce uncertainty. For example, in robot localization data from an inertial navigation system and a global positioning system GPS can be merged using filtering techniques to improve robot's navigation performance. One of the most important features of sensor fusion is that it allows to combine the information from complementary sensors, redundant sensors or even from a single sensor over a period of time. Furthermore, the advantages of using this approach are:

- Redundant information can reduce uncertainty and increase the accuracy with which features are sensed by the system.
- Multiple sensors delivering redundant information increases reliability in case of errors in a data source or when no data is available from a certain source.
- Complementary information from multiple data sources allows the perceived environment to be characterized in a way that would be impossible to perceive using only the information from each data source separately.

To merge the different emotional data sources, we have used the Kalman filter algorithm. This filtering technique is a recursive process that allows to estimate the value of the variables of interest (valence and arousal) based on knowledge of the current and previous observations, together with the description of their noise and errors. Some of the limitations of the Kalman Filter are that it can only be used for linear or linearized processes and measurement systems. However, the nature of the processed captured emotional data in terms of linear combinations of valence and arousal fit this condition. Also, the uncertainty of Kalman filter is restricted to Gaussian distribution, while other filtering techniques such as the particle filter can deal with non-Gaussian noise distribution. In any case, Kalman filtering algorithm tries to converge into correct estimations, even if the Gaussian noise parameters are poorly estimated [22].

The Kalman filter represents the system state by using two equations, where variables can be matrixes:

$$x_k = Ax_{k-1} + Bu_k + w_{k-1} \quad (3)$$

$$z_k = Hx_k + v_k \quad (4)$$

Equation (3) indicates that the data values x_k are a linear combination of its previous value, a control signal u_k and a process noise w_k-1. In our case, there is no control signal so the second term can be discarded. Equation (4) indicates that any measurement value is a linear combination of the data value and the measurement noise. While entities A, B and H are in general form matrices, in our case they can be modelled as numeric and constant as in many other signal processing problems. As described in the literature, they can be considered to be equal to 1 for simple processes [22]. Noise values w_k- and v_k are considered as Gaussian, and an approximation to their mean and standard deviation had been initially obtained from the literature on the selected emotional tools [23] – [25].

Once the system had been modelled, we have made use of a python implementation of the Kalman filter. The filtering process estimates the output at a particular state from measured data by following two steps. Firstly, the state of the system is predicted:

$$\hat{x}_{\bar{k}} = A\hat{x}_{k-1} + Bu_k \quad (5)$$

$$P_{\bar{k}} = AP_{k-1}A^T + Q \quad (6)$$

Secondly, the collected observations are incorporated once they have been corrected:

$$K_k = P_{\bar{k}}H^T(HP_{\bar{k}}H^T + R)^{-1} \quad (7)$$

$$\hat{x}_k = \hat{x}_{\bar{k}} + K_k(z_k - \widehat{Hx}_{\bar{k}}) \quad (8)$$

$$P_k = (1 - K_kH)\,P_{\bar{k}} \quad (9)$$

Where X_k is the current estimation, K_k the Kalman gain, z_k the measured value and Q is the covariance of the process noise. The process starts from an initial estimation of cero for x_0 and an error covariance matrix P_k estimated from the literature. During iteration, a prediction step is first performed, based on difference between consecutive emotion measures to compute (5) and (6). Then, as new measurements arrive (7), (8) and (9) are used to compute the estimated valence and arousal values and update the error. These estimated values are the system output at a particular state and are also used as input for the next prediction step.

IV. Results and Discussion

The system has been tested on two types of pre-recorded databases: posed and natural, as there is an important distinction between spontaneous and deliberately displayed emotions. Apart from being initiated in two different parts of the brain, but also elicited facial expressions do not look identical. Spontaneous facial expressions are characterized by synchronized, smooth, symmetrical facial muscle movements whilst posed expressions are subject to volitional real-time control and tend to be less smooth, with more variable dynamics [26].

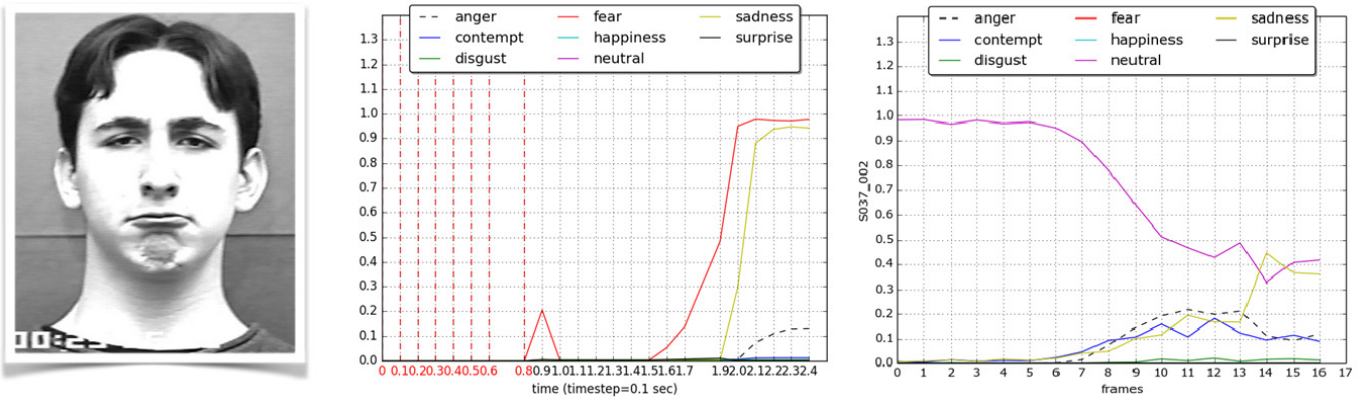


Fig. 3. Recognition results for Affectiva (center graph) and Microsoft (right graph) face emotion recognition APIs, tested on person 37, sequence 2 of the Cohn-Kanade database where the expression is labelled as sad. Affectiva was tested as live sequence whereas Microsoft was tested frame by frame. Red vertical lines indicate lost and repeated calls to Affectiva's service until the service is ready.

The selected posed databases are the Cohn-Kanade expression database [27] and the Ryerson Emotion database [28]. The Cohn-Kanade database consists in approximately 500 frontal camera image sequences from 100 subjects. Accompanying meta-data include annotation of FACS action units (AUs) (i.e., micro-expressions) [19]. However, image sequences have no sound and are only labelled in terms of action units but not emotional expressions. It is therefore necessary to translate the AU labelling for each sequence into its corresponding emotional expression. For our experiments, we have made use of the emotional labelling developed by Buenaposada et al. [29] who selected a subset of 333 sequences of 92 people from the Cohn-Kanade database and labelled them with their corresponding emotional expression. In the Ryerson Emotion database, video samples are collected from eight subjects, speaking six languages (English, Mandarin, Urdu, Punjabi, Persian, and Italian). The Ryerson Emotion database contains 720 audiovisual emotional expression samples, where subjects were provided with a list of emotional sentences and were directed to express their emotions as naturally as possible by recalling the emotional happening, which they had experienced in their lives. A frontal camera was used to record the samples in a quiet and bright environment, with a simple background.

As for natural emotional expressions, the Belfast Induced Natural Emotion Database was used. It contains recordings of mild to moderate emotionally colored responses to a series of laboratory-based emotion induction tasks [30]. The recordings are accompanied by information on self-report of emotion and intensity, continuous trace-style ratings of valence and intensity, the sex of the participant, the sex of the experimenter, and the active or passive nature of the induction task. An excerpt of the results obtained for the three databases can be found at (https://bit.ly/3OArZnp).

Facial emotion recognition APIs have been tested independently on the Cohn Kanade database to assess their performance. Overall, accuracy of both Affectiva's SDK and Microsoft's Emotional API match what is described in the literature [23] – [25]. However, Affectiva's precision (the share of correctly predicted images out of all images predicted as one category) is lower than Azure. That is, it tends to output an emotion even if not sure of what the correct emotion is. On the other hand, it has higher sensitivity (the share of correctly predicted images out of all images truly in the respective category) as Azure outputs a neutral value when unsure. Fig. 3 shows an example of this behavior. The image on the left shows the final frame (out of 16) of the sequence corresponding to person 37, sequence 2 of the Cohn-Kanade database. This sequence is composed of 16 frames that go from neutral pose to expression apex and was labelled as sad. Middle graph shows the Affectiva output, where fear and sadness are obtained as output with values close to 1. Right graph shows Microsoft output, where sad expression is only dominant during the last frames of the sequence, near the expression apex. In fact, in this case it is not very clear whether the person in the picture is sad or angry, but Affectiva still gives two high results for sadness and fear. However, there are some expressions, such as anger, that are better recognized by Affectiva and not by Microsoft. Therefore, merging these two tools can help reducing errors.

We also tested if the emotion recognition was dependent of the frame rate and if the recognition considered previous frames, as it was not clear in the case of Affectiva. In the case of Microsoft, emotion recognition is done frame by frame as stated in their documentation. As Cohn-Kanade sequences were recorded at 30 frames per second, we called the recognition services at different frame rates. Fig. 3 shows the Affectiva service being called at 10 fps. The red vertical lines correspond to lost service calls as Affectiva takes almost 1 second to initialize. Obtained results indicate that even though Affectiva accepts video as input, the recognition is little dependent of previous outputs and frame rate.

The problem with using facial recognition APIs that analyze emotions frame by frame is that there may be recognition peaks in emotion sequences even though the averaged resulting expression is correctly recognized. If analyzing micro-expressions (e.g., rising eyebrows), AU timing is related to the amount of time needed for each associated muscle to contract. On the other hand, emotional expression timing is context dependent, but it is unlikely in a real scenario to have an expression of emotion that only lasts a fraction of a second. This can lead to errors when recognizing real emotions and if the result is combined with other recognition sources. Fig. 4 shows recognition results on subject 3 sequence ha5 of the Ryerson emotion database, labelled as happiness. Left graph shows Microsoft APi output, whilst right graph shows the Kalman filter results combining Microsoft and Affectiva outputs. As it can be seen, the Kalman filter output smooths the results and removes recognition peaks.

Overall face emotion recognition tools work well on posed expressions and obtain great results on both Cohn Kanade and Ryerson databases, classifying above 80% of the sequences in the correct category. Results match those found in the literature [31], including the pattern of classifying fearful expressions as surprise or sadness, along with Microsoft's tendency to misclassify anger sadness and fear as neutral (see Fig. 3). In addition, tools that have been trained on posed and deliberate expressions fail to generalize to the complexity of expressive behavior found in real-world settings [32].

Fig. 5 (left graph) shows the results of applying just face emotion recognition tools to the analysis of natural expressions. It can be

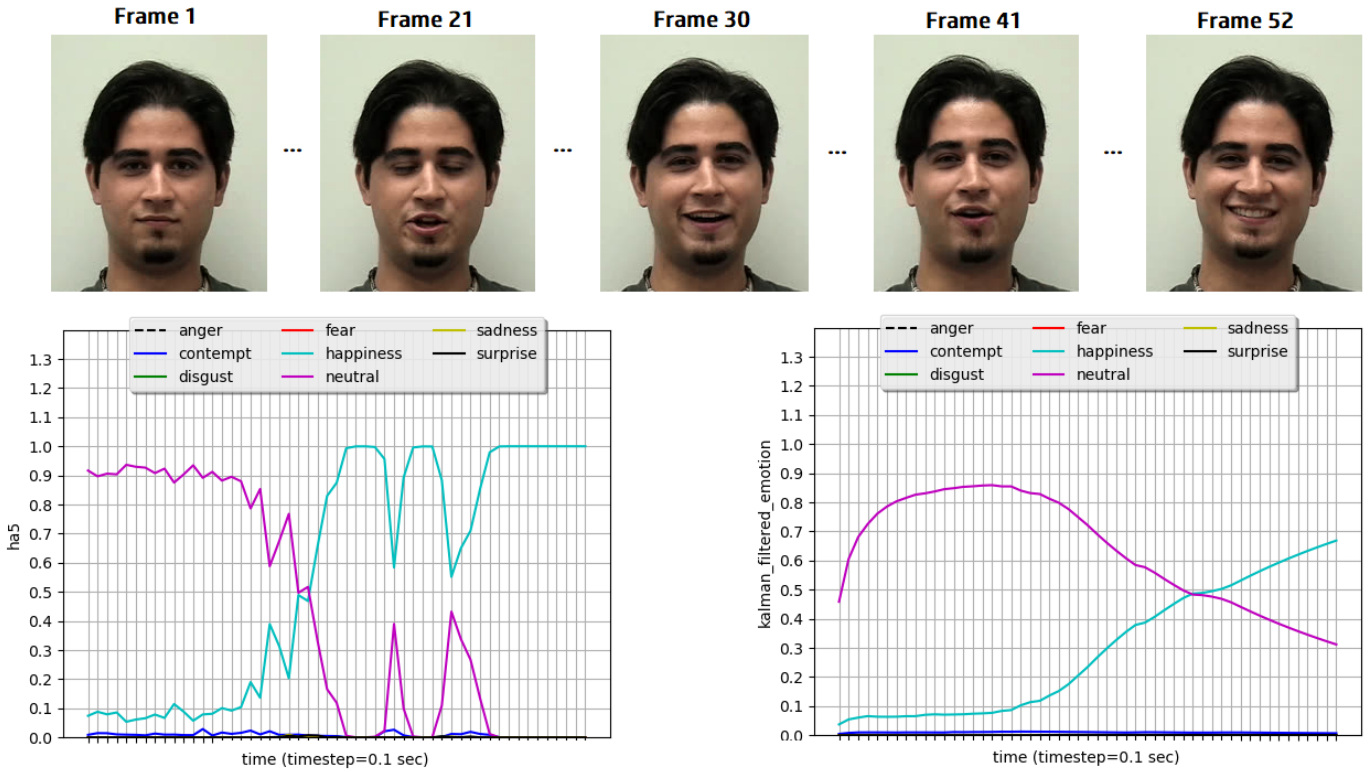


Fig. 4. Recognition results on subject 3 sequence ha5 of the Ryerson emotion database. Left graph shows Microsoft APi output, whilst right graph shous the Kalman filter results of combining Microsoft and Affectiva outputs.

observed that both tools fail to clearly output an emotion, and only in some of the initial and central frames happiness is recognized over neutral. However, images correspond to sequence 79d of the Belfast Induced Natural Emotion Database, where the task which elicited the emotion is encoded as active and social and the targeted emotion is frustration. The annotation of the sequence has negative values for valence and arousal on average, being lower in the case of arousal.

Above results show the convenience of adding other emotional sources to the recognition process, especially for the detection of spontaneous emotions. To incorporate emotion recognition from speech, captured audio is split in chunks using the SoX – Sound eXchange audio editing software. Audio file chunks can be used for prosody and text sentiment analysis. Different software has been tested for prosody analysis (Vokaturi, Beyond Verbal and openSmile). The best obtained results have been for Vokaturi, although its precision was found below 40% for both Belfast and Ryerson databases. However, as it also has very low sensitivity, prosody can be incorporated in the system provided that neutral recognition is discarded, and the error is considered during the filtering process.

On the other hand, emotion recognition from text using Microsoft Text Analytics provides improved results. As a test, we have extracted 200 reviews of New York City hotels from TripAdvisor. For each review we have saved the title, the content, and the score of the overall experience. The returned sentimental score by the text recognition tool is then compared with the number of stars assigned by the reviews' authors. The comparison is considered successful if the score is less than 40% and the assigned stars were 1 or 2 or if the score is more than 60% and the corresponding stars were 4 or 5. The reviews with 3 stars (which is supposed to be a neutral score) were not taken under account since it is unlikely that those reviews were effectively 100% neutral. Using these rules, the obtained accuracy is above 85% for positive reviews and of 89% for negative reviews.

Given this good performance, the analysis of emotions in text was therefore used to correct the deviations of the other sources of emotional recognition. That way, right graph on Fig. 5 shows the overall result of the system for recognition, where face recognition data output is translated in terms of valence and arousal and corrected as the speech recognition results are obtained.

The results show how the overall recognized emotion goes from the positive valence and arousal values obtained from the face recognition tools to negative values and close to depressed. Emotion of this sequence is self-catalogued in the Belfast database as frustration, which is associated to low levels of positive arousal and high levels of negative valence (see Fig. 1). However, we consider that obtained results are quite close to the real emotion displayed in the sequence as the person in the video speaks with a lot of laziness. In any case, obtained results are way closer to the real self-reported emotion than those obtained by the facial recognition tools.

It should be noted, however, that although obtained results for natural sequences are greatly improved, the system still fails to recognize some of the analyzed sequences from the Belfast database. These errors may be associated to two limitations of the current approach: on the one hand, by applying Kalman filtering the valence bias of the estimates obtained from the facial recognition tools can be corrected thanks to the accuracy and precision of the speech content emotion recognition measures. On the other hand, the arousal bias is only corrected by the prosodic emotion analysis, which is found to have low accuracy and precision, thus making it difficult to model. In addition, inter-subjects' variability for displaying emotions may affect the modelled error bias, as it has been built from averaged values obtained from the literature. To overcome these limitations, thanks to the modularity of the proposed architecture, additional emotional recognition sources and tools can be added in future implementations. Also, alternative approaches to the circumplex model for modelling emotions can be explored [33], as there are emotions such as fear

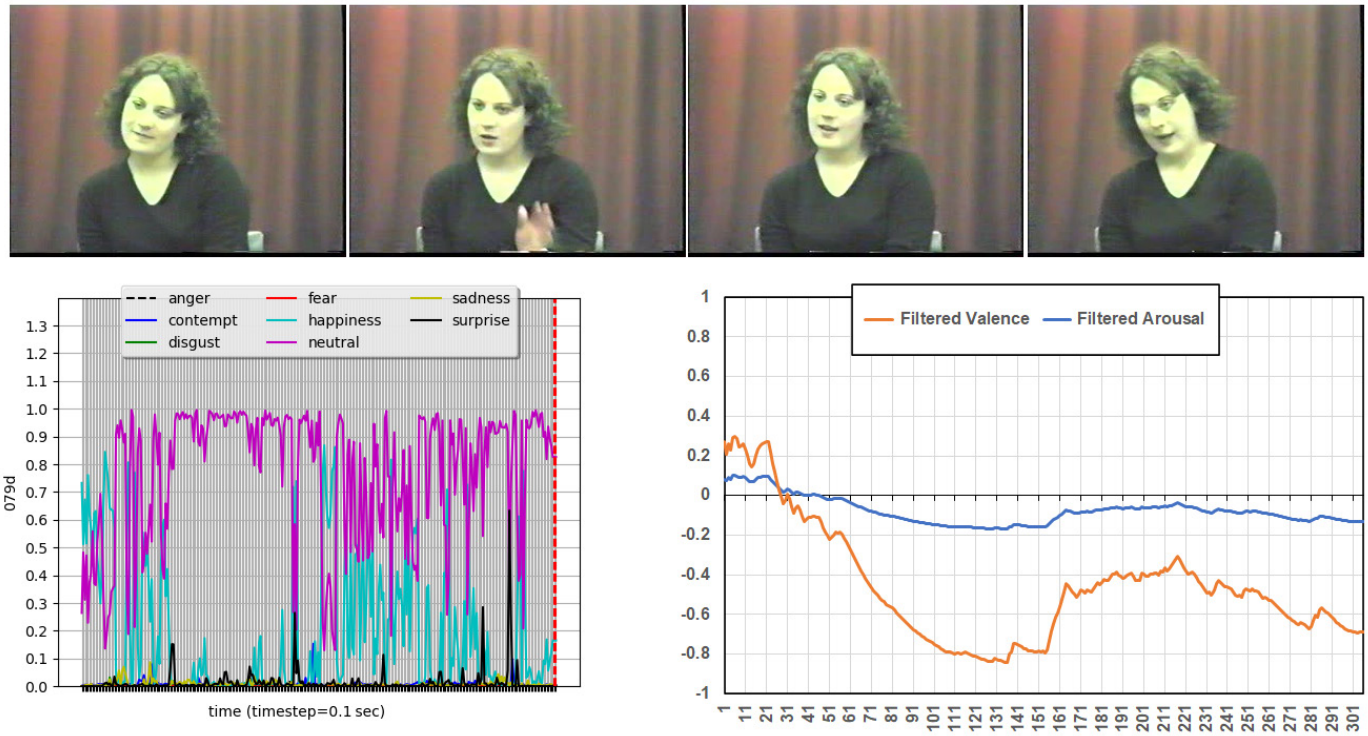


Fig. 5. Emotion recognition results corresponding to sequence 79d of the Belfast Induced Natural Emotion Database, where the task which elicited the emotion is encoded as active and social and the targeted emotion is frustration. The annotation of the sequence has negative values for valence and arousal on average, being lower in the case of arousal. Left graph shows the output obtained from facial emotion recognition tools. Right graph shows the results of the system, where face, voice and text emotion recognition sources are combined using Kalman filters for obtaining the valence and arousal of the displayed expression.

and anger (both are negative and active) that are located in the same quadrant and very close in the 2D space.

V. Conclusion

In this paper we have shown a decision-level fusion of commercially available discrete emotion recognition tools using Kalman filters. The proposed system has been modularly developed based on an evolutionary approach so to be integrated in our digital health ecosystems, and new emotional recognition sources can be easily integrated.

Obtained results show that commercially available tools such as Microsoft or Affectiva face emotion recognition APIs achieve very good recognition rates for posed expressions where no speech is involved, but their accuracy diminishes dramatically when the user communicates naturally. By fusing these tools with other recognition sources such as text analytics or prosody emotion recognition, we obtained great improvements in terms of recognizing emotions in natural databases. Moreover, with the proposed approach the output is expressed in terms of valence and arousal thus providing continuity in the emotional spectrum. This improves its potential for new applications, as allows employing existing recognition tools for more complex tasks as the ones related to health care provision

Future work should focus on adding new tools and emotional source channels to the proposed architecture and to integrate the emotion recognition system in our health ecosystem services.

Acknowledgements

This research was partially funded by the Spanish Government Ministry of Science and Innovation through the AVisSA project grant number (PID2020-118345RB-I00).

References

[1] A. García-Holgado, S. Marcos-Pablos, F.J. García-Peñalvo, "A Model to Define an eHealth Technological Ecosystem for Caregivers", in *New Knowledge in Information Systems and Technologies*, Á. Rocha, H. Adeli, L. P. Reis, and S. Costanzo, Eds. Springer International Publishing, 2019, pp. 422–432, https://doi.org/10.1007/978-3-030-16187-3_41.

[2] S. Marcos-Pablos, A. García-Holgado, F.J. García-Peñalvo, "Modelling the business structure of a digital health ecosystem", in *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2019, pp. 838–846, https://doi.org/10.1145/3362789.3362949.

[3] S. Marcos-Pablos, F.J. García-Peñalvo, "Emotional Intelligence in Robotics: A Scoping Review", In *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence*, J. F. de Paz Santana, D. H. de la Iglesia, & A. J. López Rivero, Eds. Springer International Publishing, 2022, pp. 66–75, https://doi.org/10.1007/978-3-030-87687-6_7.

[4] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion", *Information Fusion*, 2017, vol. 37, pp. 98–125, https://doi.org/10.1016/j.inffus.2017.02.003.

[5] M.G. Huddar, S.S. Sannakki, and V.S. Rajpurohit, "Attention-based Multimodal Sentiment Analysis and Emotion Detection in Conversation using RNN", International Journal of Interactive Multimedia and Artificial Intelligence, 2021, vol. 6, no. 6, pp. 112-121, http://doi.org/10.9781/ijimai.2020.07.004.

[6] H. Daus and M. Backenstrass, "Feasibility and Acceptability of a Mobile-Based Emotion Recognition Approach for Bipolar Disorder", International Journal of Interactive Multimedia and Artificial Intelligence, 2021, vol. 7, no. 2, pp. 7-14, http://doi.org/10.9781/ijimai.2021.08.015.

[7] M. Magdin, D. Držík, J. Reichel, and S. Koprda, "The Possibilities of Classification of Emotional States Based on User Behavioral Characteristics", International Journal of Interactive Multimedia and Artificial Intelligence, 2020, vol. 6(Regular Issue), no. 2, pp. 97-104, http://doi.org/10.9781/ijimai.2020.11.010.

[8] S. Kirrane, "Intelligent software web agents: A gap analysis", *Journal*

of Web Semantics, 2021, vol. 71, 100659, https://doi.org/10.1016/j.websem.2021.100659.

[9] W. Brenner, R. Zarnekow, and H. Wittig, "Intelligent Software Agents: Foundations and Applications", Springer-Verlag Berlin, 1998, https://doi.org/10.1007/978-3-642-80484-7.

[10] P. Salovey, and J. D. Mayer, "Emotional Intelligence", *Imagination, Cognition and Personality*, 1990, vol. 9, no. 3, pp. 185–211, https://doi.org/10.2190/DUGG-P24E-52WK-6CDG.

[11] W. Ickes, "Empathic Accuracy", *Journal of Personality*, 1993, vol. 61, no. 4, pp. 587–610, https://doi.org/10.1111/j.1467-6494.1993.tb00783.x.

[12] E. van der Kruk, M. M. Reijne, "Accuracy of human motion capture systems for sport applications; state-of-the-art review", European Journal of Sport Science, 2018, vol. 18, no. 6, pp. 806–819, https://doi.org/10.1080/17461391.2018.1463397.

[13] L. Shu, J. Xie, M. Yang, Z. Li, D. Liao, X. Xu, and X. Yang, "A Review of Emotion Recognition Using Physiological Signals", *Sensors*, 2018, vol. 18, no. 7, 2074, https://doi.org/10.3390/s18072074.

[14] M.L. Rohlfing, D.P. Buckley, J. Piraquive, C.E. Stepp, and L.F. Tracy, "Hey Siri: How Effective are Common Voice Recognition Systems at Recognizing Dysphonic Voices?", *Laryngoscope*, 2021, vol. 131, no. 7, pp. 1599-1607, https://doi.org/10.1002/lary.29082.

[15] N. Samadiani, G. Huang, B. Cai, W. Luo, C.H. Chi, Y. Xiang, and J. He, "A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data", *Sensors*, 2019, vol. 19, no. 8, https://doi.org/10.3390/s19081863.

[16] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments", *Information Fusion*, 2201, vol. 68, pp. 46–53, https://doi.org/10.1016/j.inffus.2020.10.011.

[17] D.C. Rubin and J.M. Talarico, "A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words", *Memory*, 2009, vol. 17, no. 8, pp. 802–808, https://doi.org/10.1080/09658210903130764.

[18] S. Marcos-Pablos, F.J. García-Peñalvo, and A. Vázquez-Ingelmo, "Emotional AI in Healthcare: A pilot architecture proposal to merge emotion recognition tools" in *Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2021, pp. 342–349, https://doi.org/10.1145/3486011.3486472.

[19] P. Ekman and E.L. Rosenberg, "What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)", *Oxford University Press*, 2005, https://doi.org/10.1093/acprof:oso/9780195179644.001.0001.

[20] G. Paltoglou, M. Thelwall, "Seeing Stars of Valence and Arousal in Blog Posts", in *IEEE Transactions on Affective Computing*, 2013, vol. 4, no. 1, pp. 116–123, https://doi.org/10.1109/T-AFFC.2012.36.

[21] M. Olszanowski, G. Pochwatko, K. Kuklinski, M. Scibor-Rylski, P. Lewinski, and R.K. Ohme, "Warsaw set of emotional facial expression pictures: A validation study of facial display photographs" *Frontiers in Psychology*, 2015, vol. 5, https://www.frontiersin.org/article/10.3389/fpsyg.2014.01516.

[22] B. Ristic, S. Arulampalam, and N. Gordon, "Beyond the Kalman Filter: Particle Filters for Tracking Applications", *Artech House*, 2003.

[23] A. Bhattacharjee, T. Pias, M. Ahmad, and A. Rahman,"On the Performance Analysis of APIs Recognizing Emotions from Video Images of Facial Expressions", *17th IEEE International Conference on Machine Learning and Applications*, 2018, pp. 223–230, https://doi.org/10.1109/ICMLA.2018.00040.

[24] A. Mathur, A. Isopoussu, F. Kawsar, R. Smith, N.D. Lane, and N. Berthouze, "On Robustness of Cloud Speech APIs: An Early Characterization", in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018, pp. 1409–1413, https://doi.org/10.1145/3267305.3267505.

[25] S.R. Khanal, J. Barroso, N. Lopes, J. Sampaio, and V. Filipe, "Performance analysis of Microsoft's and Google's Emotion Recognition API using pose-invariant faces", in *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion*, 2018, pp. 172–178, https://doi.org/10.1145/3218585.3224223.

[26] W.E. Rinn, "The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions", *Psychological Bulletin*, 1984, vol. 95, no. 1, pp. 52–77.

[27] T. Kanade, J.F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis", in *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 46–53, https://doi.org/10.1109/AFGR.2000.840611.

[28] Ryerson Emotion Database. (n.d.). Retrieved April 23, 2022, from https://www.kaggle.com/datasets/ryersonmultimedialab/ryerson-emotion-database

[29] J.M. Buenaposada, E. Muñoz, and L. Baumela, "Recognising facial expressions in video sequences", *Pattern Analysis and Applications*, 2008, vol. 11, no. 1, pp. 101-116.

[30] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The Belfast Induced Natural Emotion Database", *IEEE Transactions on Affective Computing*, 2012, vol. 3, no. 1, pp. 32–41, https://doi.org/10.1109/T-AFFC.2011.26.

[31] T. Küntzler, T.T.A. Höfling, and G.W. Alpers, "Automatic Facial Expression Recognition in Standardized and Non-standardized Emotional Expressions", *Frontiers in Psychology*, 2021, vol. 12, https://www.frontiersin.org/article/10.3389/fpsyg.2021.627561.

[32] "A New Video Based Emotions Analysis System (VEMOS): An Efficient Solution Compared to iMotions Affectiva Analysis Software". (n.d.). ASTES Journal. Retrieved April 25, 2022, from https://astesj.com/v06/i02/p114/.

[33] Z. Kowalczuk and M. Czubenko, "Computational Approaches to Modeling Artificial Emotion – An Overview of the Proposed Solutions", Frontiers in Robotics and AI, 2016, vol. 3, https://doi.org/10.3389/frobt.2016.00021.

Samuel Marcos-Pablos

He received a Telecommunication Engineer's Degree in 2006, a M.Eng. in robotics in 2009, and a Ph.D. in robotics in 2011 from the University of Valladolid (Spain). He has worked as a researcher at CARTIF's Robotics and Computer Vision Division from 2007 - 2018, where he combined theoretical and field work in the research and development of projects in the area of Social and Service robotics and computer vision. He is currently with the GRIAL research group, and focuses his efforts in the development of ecosystems for the health sector and teaching. Among others, he has authored papers for the journals of Interacting With Computers or Sensors MDPI, as well as conferences such as the IEEE International Conference on Intelligent Robots and Systems and the IEEE International Conference on Robotics and Automation.

Fernando Lobato Alejano

Fernando Lobato Alejano is a Ph.D. in Computer Engineering from the Pontifical University of Salamanca, a Technical Engineer in Computer Systems and has a Degree in Computer Engineering from the Catholic University of Murcia. He also has Master's Degree in teaching, specializing in technology. He is the author of different book chapters and has a multitude of intellectual property registrations, as well as a utility model and a patent in progress. He has been awarded in the cross-border competition for market-oriented prototypes (Prototransfer/Inespo 2018) and currently works as researcher and Professor at the Pontifical University of Salamanca in the areas of Computing Engineering and Management of Technology-Based Companies.

Francisco José García-Peñalvo

He received the degrees in computing from the University of Salamanca and the University of Valladolid, and a Ph.D. from the University of Salamanca (USAL). He is Full Professor of the Computer Science Department at the University of Salamanca. In addition, he is a Distinguished Professor of the School of Humanities and Education of the Tecnológico de Monterrey, Mexico. Since 2006 he is the head of the GRIAL Research Group GRIAL. He is head of the Consolidated Research Unit of the Junta de Castilla y León (UIC 81). He was Vice-dean of Innovation and New Technologies of the Faculty of Sciences of the USAL between 2004 and 2007 and Vice-Chancellor of Technological Innovation of this University between 2007 and 2009. He is currently the Coordinator of the PhD Programme in Education in the Knowledge Society at USAL. He is a member of IEEE (Education Society and Computer Society) and ACM.

LA UNIVERSIDAD EN INTERNET

Rectorado
Avenida de la Paz, 137
26006 Logroño (La Rioja)
t (+34) 941 21 02 11

www.unir.net