Pattern Recognition 122 (2022) 108283

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

SetMargin loss applied to deep keystroke biometrics with circle packing interpretation

Aythami Morales*, Julian Fierrez, Alejandro Acien, Ruben Tolosana, Ignacio Serna

School of Engineering, Universidad Autonoma de Madrid, Spain

ARTICLE INFO

Article history:
Received 23 April 2021
Revised 13 August 2021
Accepted 28 August 2021
Available online 29 August 2021

Keywords:
Keystroke biometrics
Circle packing
Deep learning
DML

ABSTRACT

This work presents a new deep learning approach for keystroke biometrics based on a novel Distance Metric Learning method (DML). DML maps input data into a learned representation space that reveals a "semantic" structure based on distances. In this work, we propose a novel DML method specifically designed to address the challenges associated to free-text keystroke identification where the classes used in learning and inference are disjoint. The proposed SetMargin Loss (SM-L) extends traditional DML approaches with a learning process guided by pairs of sets instead of pairs of samples, as done traditionally. The proposed learning strategy allows to enlarge inter-class distances while maintaining the intra-class structure of keystroke dynamics. We analyze the resulting representation space using the mathematical problem known as Circle Packing, which provides neighbourhood structures with a theoretical maximum inter-class distance. We finally prove experimentally the effectiveness of the proposed approach on a challenging task: keystroke biometric identification over a large set of 78,000 subjects. Our method achieves state-of-the-art accuracy on a comparison performed with the best existing approaches.

1. Introduction

In a global society migrating from physical services to digital platforms, identity management becomes critical. However, traditional physical user authentication cannot be directly applied in digital services. Keystroke biometric recognition enables the identification of users based on their typing behavior. Keystroke biometric systems are commonly placed into two categories: *fixed-text*, characterized by a prefixed keystroke sequence typed by the user (e.g. passwords), and *free-text*, characterized by arbitrary keystroke sequences (e.g. emails or transcriptions). Free-text systems must therefore consider different text content between training and testing, including typing errors.

Keystroke dynamics authentication literature has been predominantly focused on verification tasks in fixed-text scenarios. Approaches based on statistical models (e.g. Hidden Markov Models) [1], Manhattan distances [2], sample alignment (e.g. Dynamic Time Warping) [3], and digraphs [4] have achieved competitive results in fixed-text verification [5]. The performance in free-text scenarios remained far from those reached in the fixed-text verification approaches during the last decade. Partially Observable Hidden Markov Models were employed in [6] for free-text keystroke verification obtaining a competitive accuracy. More recently, the availability of large scale databases with millions of keystroke samples has allowed training deep models with very competitive performances in free-text scenarios [7]. The architecture proposed in [7], called TypeNet, was trained using a Contrastive Loss function with performances six times better than previous approaches based on traditional statistical methods [6,8]. Our purpose in the present paper is to improve further the state-of-the-art results of deep keystroke biometrics by introducing a new loss function expected to be also useful in other challenging recognition problems.

There are two main research lines to define such loss functions: *i)* approaches based on Distance Metric Learning (DML) such as Contrastive Loss [9], Triplet Loss [10], and their variants [11,12]; and *ii)* with multi-class classifiers based on Softmax Loss functions and its variants [13–15]. Both research lines present advantages and disadvantages.

In the first line of work (DML), the core idea is to train a function that maps input data into a new feature space where simple distances can serve to analyze and exploit the "semantic" structure of the input space [9]. A DML approach serves to define a neighbourhood structure in the feature space based on a relationship

* Corresponding author.
E-mail addresses: aythami.morales@uam.es (A. Morales), julian.fierrez@uam.es (J. Fierrez), alejandro.acien@uam.es (A. Acien), ruben.tolosana@uam.es (R. Tolosana), ignacio.serna@uam.es (I. Serna).

https://doi.org/10.1016/j.patcog.2021.108283

between intra-class (between samples from the same class) and inter-class distances (between samples from different classes). In an ideal feature space, samples from the same class will remain "near" and samples from different classes will be pushed "far". Near and far can be defined based on simple distances like Euclidean. Noteworthy, most of the DML approaches in the literature are based on learning processes based on pairs of samples [9,10,13].

In the second line of work (i.e., using multi-class classifiers), there are some limitations stemmed from using classifiers. A classification algorithm is mostly associated to categorization tasks, but it can be used to tackle other representation learning problems. One example is the use of classification algorithms as feature extractors where models are trained for classification, and the outputs (usually the last layers of a deep network) are employed as features for other tasks [15–17]. However, using classification algorithms to learn discriminatory feature spaces exhibits limitations. On the one hand, the feature space learned might not be suitable for classes not seen during learning. On the other hand, the error propagated during learning is based on a scalar prediction (i.e., a label), which is a simplification of the whole problem at hand defined by intra- and inter-class neighborhood structures [9]. To address these problems, some authors have proposed methods based on the joint supervision of Softmax Loss and DML to improve the discrimination power of the feature learned spaces [13,15].

In the present work, we propose a novel DML approach (SetMargin Loss, SM-L) specifically designed to address the challenges associated to free-text keystroke identification where classes used in learning and inference are disjoint. The final aim is to identify the membership of the input data to a class unseen during learning. SM-L extends traditional DML approaches with a learning process guided by pairs of sets, which allows to enlarge inter-class distances while maintaining the intra-class structure.

We will analyze the learned feature space generated with SM-L in comparison with other popular loss functions using the mathematical problem known as Circle Packing. The solution to the Circle Packing problem is a neighbourhood structure that guarantees a theoretical maximum inter-class distance. We propose using this method to gain understanding in the feature spaces obtained by DML approaches. Finally, we will prove experimentally the effectiveness of our proposed SM-L on a challenging task: keystroke biometric identification [7]. The proposed approach outperforms other popular loss functions in this problem.

In summary, the contributions of this work are:

- A new keystroke identification/verification method based on a novel loss function called SetMargin Loss (SM-L).
- We introduce the Circle Packing problem as a novel way to gain insights into learned feature spaces.
- We experiment with the proposed SM-L on a challenging open-set keystroke biometric identification/verification problem over 78,000 subjects, achieving state-of-the-art performance superior to related methods.

The rest of the paper is organized as follows. Next section summarizes the most popular DML approaches and presents the Circle Packing Problem as a way to analyze feature spaces. The third section describes the proposed SetMargin Loss and the fourth section presents the experiments and results. Finally, the work finish with the conclusions.

2. Distance metric learning: loss functions

The objective of metric learning is to generate distances d between input data pairs either from the same or different classes (positive and negative pairs, respectively) useful for a certain task where a component of the distance is based on a learned model related to the task at hand. These distances d can be defined, e.g., as Euclidean distances:

$$d(\mathbf{x}^i, \mathbf{x}^j) = \left\| \mathbf{f}(\mathbf{x}^i|\mathbf{w}) - \mathbf{f}(\mathbf{x}^j|\mathbf{w}) \right\| \tag{1}$$

where $\mathbf{w}$ are the weights of a model (typically a neural network), and $\mathbf{f}(\mathbf{x}^i|\mathbf{w})$, $\mathbf{f}(\mathbf{x}^j|\mathbf{w})$ are the model outputs (embedding vectors) for the inputs $\mathbf{x}^i$ and $\mathbf{x}^j$, respectively.

There are several metric learning approaches in the literature [9–11,13]. Among these approaches, the *Contrastive Loss* function [9] is a popular example of DML technique with a history of success in many applications [7,18,19]. Let $\mathbf{x}^i$ and $\mathbf{x}^j$ each be a sample that together form a pair which is provided as input to a Siamese Neural Network [18] with shared weights $\mathbf{w}$. The loss function $\mathcal{L}_{CL}$ is defined as follows:

$$\mathcal{L}_{CL} = (1 - L_{ij}) \frac{d^2(\mathbf{x}^i, \mathbf{x}^j)}{2} + L_{ij} \frac{\max^2 \left\{0, \alpha - d(\mathbf{x}^i, \mathbf{x}^j)\right\}}{2} \tag{2}$$

where L_{ij} is a label associated with each pair that is set to 0 for positive pairs and 1 for negative ones, and $\alpha \geq 0$ is a margin. As we can see, the Contrastive Loss learns intra- and inter-class distances in separate operations defined by L_{ij}.

The *Triplet Loss* function [10,20] appeared as a function to learn from positive and negative comparisons at the same time. A triplet is defined by three samples known as Anchor, Positive, and Negative. Anchor ($\mathbf{x}^i_A$) and Positive ($\mathbf{x}^i_P$) are samples from the same class i, while Negative ($\mathbf{x}^j_N$) is a sample from a different class j. The Triplet loss function is defined as follows:

$$\mathcal{L}_{TL} = \max \left\{0, d^2(\mathbf{x}^i_A, \mathbf{x}^i_P) - d^2(\mathbf{x}^i_A, \mathbf{x}^j_N) + \alpha\right\} \tag{3}$$

where α is a margin between positive and negative pairs. In comparison with Contrastive Loss, Triplet Loss is capable of learning intra-class and inter-class structures in a unique operation (removing the label L_{ij}).

There are other metric learning methods designed to guide the learning process in different ways. The *Center Loss* was proposed to minimize the intra-class distances of the deep features [13]. Center Loss combines the traditional Softmax with a loss function aimed to reduce the distance of feature vectors to an average feature vector calculated for each class (i.e., centroid). Similarly, the *Magnet Loss* [21] introduces a learning approach inspired in clustering techniques where the loss function depends on cluster optimization instead of traditional sample classification. *N-Pair Loss* is an extension of Triplet Loss to several negative samples [11], where triplets are conformed using one positive sample and N multiple negative examples. More recently, researchers have proposed the *Angular Softmax Loss* [15] to improve face recognition performance. These methods improve the traditional Softmax Loss function by incorporating an angular learning objective.

2.1. Circle packing and learned feature spaces

In this section we introduce the Circle Packing problem to gain insights into the neighborhood structures learned in feature spaces. A packing of circles as defined in [22] is a collection (finite or infinite) of circles $P = \{C_1, \ldots, C_N\}$ on a given (Riemann) surface S with disjoint interiors (i.e., distinct circles in P may be tangent, but cannot overlap). This is the general definition. In our specific case, the Riemann surface is a closed circular region $D \subset \mathbb{R}^2$ with the standard Euclidean metric, the packing P is finite, and all circles C_i have unit-radius. Thus, our objective can be defined as finding a Circle Packing P such that the minimum distance between circles is maximized. The result is a structure that minimizes the area outside the unit circles, and therefore the radius R^N of the outer circle D. The solution depends on the number of circles N and the analytical demonstration varies depending on N (i.e., there is not a unique way to solve the problem for different N). Fig. 1 shows the solutions from $N = 8$ to $N = 13$ proved by Melissen [23], Fodor [24],

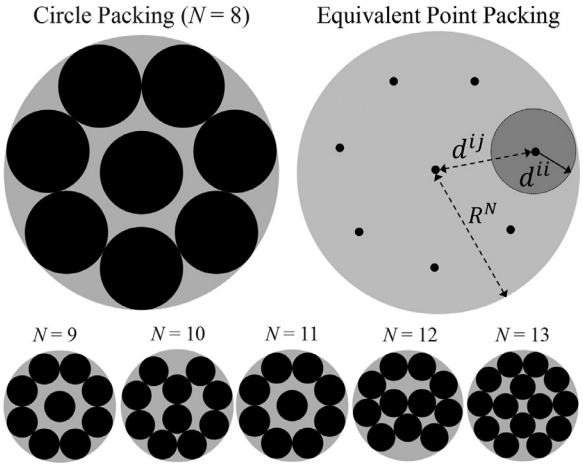

Fig. 1. Circle Packing solution from $N = 8$ to $N = 13$ and equivalent Point Packing problem for $N = 8$. d^{ij} and d^{ii} represent the inter- and intra-class distances respectively. R^N is the minimum radius of the circular region D that contains the unit circles.

25]. Note that, as also shown in Fig. 1, the Circle Packing problem can be transformed into a Point Packing problem by replacing circles by their centers. This mathematical problem can be seen as an optimization task, and many computational approaches have been proposed to solve it [26,27]. To the best of our knowledge, this is the first time that Circle Packing is used to analyze learned feature spaces.

The solution to the Circle Packing problem maximizes the distances between circle centers d^{ij} subject to fixed radii. Assuming that each circle is a class in our feature space, the distance d^{ij} represents the inter-class distance (i.e., distance between samples from different classes) while d^{ii} represents the intra-class distance (i.e., distance of samples from the same class). This formulation is specially useful in open-set classification problems, where we seek a feature space that: *i)* maximizes inter-class distances; and *ii)* minimizes intra-class distances. To achieve the maximum inter-class distance, it is necessary to distribute classes along all the space. It is important to keep in mind that in an open-set scenario the classes used during training are different to those used for testing.

We have conducted a toy example to visualize the feature space obtained by different loss functions and its similarity to the Circle Packing optimal solution. To this end, we use a subset of the "Quick, Draw!" dataset [28]. This dataset comprises 50 million drawings across 345 different categories. This database is interesting because of the large intra-class variability in the different classes (e.g. there are hundreds of different ways to draw a plane). Each drawing is converted to a 28×28 grey scale image. In order to visualize the feature space learned by typical deep models, we train a Convolutional Neural Network (CNN) inspired in the popular VGG architecture [29] and composed of: two Convolutional layers (32 and 64 units, 3×3 filter bank size, ReLU activation), 2D Maxpooling layer, Dense layer (128 units, ReLU activation), Dense layer mapping the features into a 2D space (2 units, Linear activation), and Output layer (13 units, Softmax activation).

We use 100 images from the first 8 classes of the "Quick, Draw!" dataset to train a classifier (batch size = 32, Adam optimizer, learning rate = 0.01). Fig. 2 shows an example of how different loss functions define the feature space (for Contrastive Loss and Triplet Loss, we have removed the final output layer of the model). The feature spaces are generated plotting the output of the 2 units layer included in the CNN model. The feature space obtained by Softmax Loss has an intrinsic angular distribution as it is expected [15]. The angular distribution obtained by Softmax is far from the Circle Packing solution, but it is not necessarily a wrong solution. Recent approaches based on softmax angular margin losses have achieved state-of-the-art performances in Face Recognition problems [15]. Contrastive Loss [9] maximizes the inter-class distances but fails to exploit all the available space. The feature space obtained by this loss function is similar to the one obtained by other loss functions that maximize inter-class distances in a joint supervision with Softmax [13]. Finally, Triplet Loss [10] shows a feature space that perfectly matches the optimal Circle Packing solution (see Fig. 2). The feature space generated by Triplet Loss approximates the theoretical maximum inter-class distance in a feature space divided into 8 circular regions.

2.2. Limitations of the circle packing solution

Besides the several similarities between the optimal solution to the Circle Packing problem and the feature space generated with a specific learning strategy, we have to consider some limitations in this comparison:

- The results obtained by DML approaches are usually characterized by a separable Euclidean space. For approaches based on Euclidean spaces, the Circle Packing framework can be used to find a theoretical maximum inter-class distance. However, defining each class region as a unit circle is not necessarily the best approach for all problems. The distance d^{ii} can vary between classes (i.e., intra-class variability depends on the class) and we are projecting into a 2-dimensional feature space assuming low correlation between features. In the present paper we simplify the problem assuming unit circles, but Circle Packing can be extended to circles with different area [30] or ellipses with different shape [31].
- Most of the learned feature spaces are characterized by more than 2 dimensions. Extensions can be made to higher dimensions. In 3 dimensions the equivalent problem is known as Sphere Packing; and Hypersphere Packing in higher dimensions.
- In the present paper, we will show how these spaces are suitable for open-set classification problems. Nonetheless, a feature space defined by the Circle Packing solution is not necessarily the best solution for all machine learning problems, e.g., that solution does not guarantee good generalization properties to unseen data.

3. Proposed method: Setmargin loss (SM-L)

The main challenges associated to free-text keystroke identification are: *i)* large intra-class variability (i.e., the typing behavior of a given subject may vary occasionally); *ii)* low inter-class variability (i.e., the typing behavior from different subjects might be similar); *iii)* large number of classes (one per subject which can easily scale to several thousands); *iv)* free-text scenario (i.e., identification must be performed with independence of the text typed); and *v)* small to moderate number of samples per class available to model the problem (i.e., 15 samples per class in our case). These challenges associated to keystroke biometric identification have made to fail recent machine learning approaches in this task for identifying a large number of subjects [7].

Here we propose to overcome these challenges with an extension of Tripet Loss. In comparison with Contrastive Loss, Triplet Loss allows to model the relationship between positive and negative samples in a unique operation (see Fig. 3 and Eq. (3)). Both methods were developed to learn from comparisons made with pairs of samples $d(\mathbf{x}^i, \mathbf{x}^j)$. A learning process guided by pairs of samples may not be adequate for the possibly complex intra-class relationships between samples of the same class. With our pro-

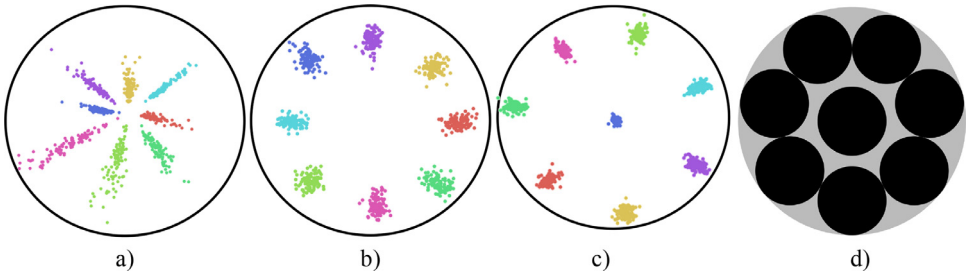

Fig. 2. 2D feature spaces ($N = 8$) learned by: (a) Softmax Loss, (b) Contrastive Loss, (c) Triplet Loss, and (d) Circle Packing optimal solution for $N = 8$.

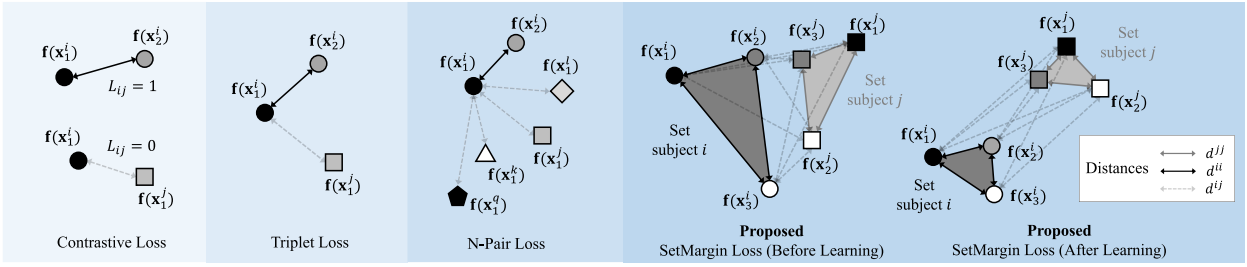

Fig. 3. Sets of distances considered in popular deep learning loss functions: Contrastive Loss, Triplet Loss, N-Pair Loss, and the proposed SetMargin Loss (SM-L) for a pair of sets with $G = 3$ samples per set. Shapes represent different classes while color indicates different samples for the same class.

posed SetMargin Loss (SM-L) we propose to extend this learning strategy to pairs of sets instead of pairs of samples. This learning strategy allows to capture better intra-class dependencies while enlarging the inter-class differences in the feature space (see Fig. 3).

In practice, there are different ways to transform a sample-pair based learning into a sample-set learning process. We propose to evaluate two different implementations of our idea of set distances: *SetMargin Contrastive Loss (SM-CL)* and *SetMargin Triplet Loss (SM-TL)*.

Let $\{\mathbf{x}_k^i\}_{k=1,\dots,G^i}$ and $\{\mathbf{x}_q^j\}_{q=1,\dots,G^j}$ be a pair of sets provided as input to the model. The *SetMargin Contrastive Loss (SM-CL)* proposed in this work is an extension of Eq. (2) defined as follows:

$$\mathcal{L}_{SM-CL} = \sum_{k=1}^{G^i}\sum_{q=k+1}^{G^i}\frac{d^2(\mathbf{x}_k^i,\mathbf{x}_q^i)}{2} + \beta\sum_{k=1}^{G^i}\sum_{q=1}^{G^j}\frac{\max^2\left\{0,\alpha - d(\mathbf{x}_k^i,\mathbf{x}_q^j)\right\}}{2} \quad (4)$$

where α is a margin, $d(\cdot)$ is the Euclidean distance defined in Eq. (1) and β is a constant that serves to weight the intra-class and inter-class distances. In our experiments $G^i = G^j = G$ is the number of samples per class and $\beta = 2G$ is proportional to the number of learning samples per class.

The *SetMargin Triplet Loss (SM-TL)* proposed in this work is an extension of traditional Triplet Loss (see Eq. (3)) to learn from pairs of sets instead of pair of samples. This extension adds the context of the set to the learning process resulting in a large-margin representation capable of improving the distance between classes. The loss function is calculated as follows:

$$\mathcal{L}_{SM-TL} = \sum_{k=1}^{G^i}\sum_{q=k+1}^{G^i}\sum_{l=1}^{G^j}(\max\left\{0, d^2(\mathbf{x}_k^i,\mathbf{x}_q^i) - d^2(\mathbf{x}_k^i,\mathbf{x}_l^j) + \alpha\right\} + \max\left\{0, d^2(\mathbf{x}_k^j,\mathbf{x}_q^j) - d^2(\mathbf{x}_k^j,\mathbf{x}_l^i) + \alpha\right\}) \quad (5)$$

Note that we assume $G^i = G^j = G$ but the method can be directly extended to problems where $G^i \neq G^j$. The margin α is equal to 1.5 in all our experiments. The literature has shown that triplet selection can significantly improve the quality of the learned spaces [32,33]. Our approach does not include a direct selection of triplets. Nonetheless, the *max* function included in Eq. (5) is used to reduce the impact of "easy triplets" (i.e., $d^2(\mathbf{x}_k^i,\mathbf{x}_l^j) > d^2(\mathbf{x}_k^i,\mathbf{x}_q^i) + \alpha$).

3.1. Intuition of the learning process

Without loss of generality let's assume a template composed of 3 samples ($G = 3$). $\mathcal{T}^i$ is the triangle whose vertices are each of the three embeddings of the template of subject i and d^{ij} is the distance between barycenters from $\mathcal{T}^i$ to $\mathcal{T}^j$. The *SetMargin Loss* minimizes the areas of $\mathcal{T}^i$ and $\mathcal{T}^j$ while maximizing the distance d^{ij} (close to α). Incorporating the template geometry into the learning objective we enrich the learned space generation process and this will result in better embedding representations (see Fig. 4).

The main characteristics of the proposed SM-L metric learning are: *i)* it maximizes the inter-class distance while preserving intra-class compactness through batches composed by pairs of sets instead of pairs of samples; *ii)* it produces highly discriminative feature spaces adequate for open-set classification problems where inference is conducted on samples of classes unseen in the learning process, and therefore the feature space is constructed over unseen class relationships; and *iii)* it is able to learn highly discriminative feature spaces from a limited number of samples per class. The number of possible set combinations is very high and the set generation acts as a data augmentation technique.

3.2. Comparison with other loss functions

Fig. 4 shows the feature space learned by different loss functions using the toy example presented in previous sections. The figure depicts the feature spaces for: Softmax Loss, ArcFace Loss [15], Contrastive Loss [9], Triplet Loss [10], N-Pair Loss [11], and our SetMargin Loss in the two proposed implementations: SM-CL and SM-TL. The angular distribution of classes observed for the Softmax Loss function is enhanced by the ArcFace Loss function [15]. Contrastive Loss shows a feature space with a common pattern where all classes are distributed in the exterior regions and

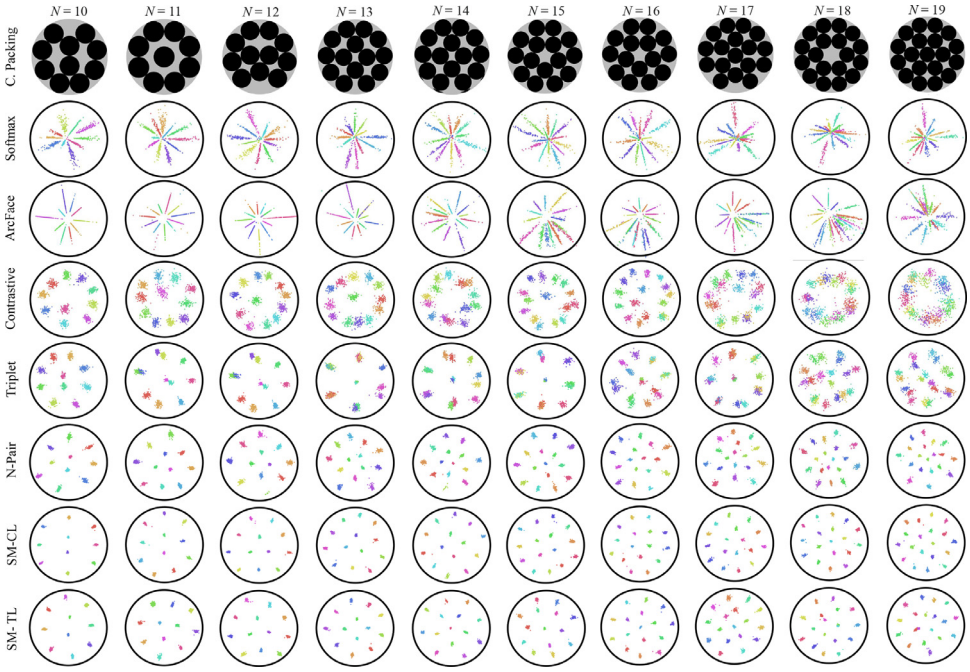

Fig. 4. Circle Packing problem solutions from $N = 10$ to $N = 19$ and feature spaces learned by different loss functions.

one class is located at center. These spaces are similar to those obtained by other loss functions such as Center Loss [13]. For large number of classes, this type of distribution is highly inefficient. Triplet Loss tends to create spaces with structures similar to those of the Circle Packing solution, but fails with N greater than 10. N-Pair Loss is based on a learning process guided by inter-class comparisons. Thus, N-Pair Loss improves the margin between classes with respect to Triplet Loss. The proposed SetMargin Loss shows feature spaces very similar to those obtained by the optimal solution to the Circle Packing problem. The feature space obtained by our method guarantees a good separation between classes, close to the theoretical maximum.

We propose two quantitative metrics to evaluate the distribution of classes in the learned spaces: minimum distance between centroids (δ), and intra-cluster dispersion (ρ). The minimum distance between centroids δ is calculated as:

$$\delta = \frac{1}{N}\sum_{i=1}^{N} \min_{j} ||\mathbf{c}^i - \mathbf{c}^j||, (i \neq j) \tag{6}$$

where N is the number of classes and $\{\mathbf{c}^i, \mathbf{c}^j\}$ are the centroids of the embeddings of the classes i and j. The intra-cluster dispersion ρ is calculated as:

$$\rho = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{L^i}\sum_{k=1}^{L^i}\|\mathbf{c}^i - \mathbf{f}(\mathbf{x}_k^i)\|\right) \tag{7}$$

where L^i is the total number of data points (i.e., number of samples) of the class $i (L^i \geq G^i)$ and $\mathbf{f}(\mathbf{x}_k^i)$ is the embedding vector k of the class i. The maximum δ for a given N can be calculated using the Circle Packing solution as:

$$\delta_{max}^N = \frac{1}{2R^N - 1} \tag{8}$$

where R^N is the minimum radius of the circular region D that contains the N unit circles. In a similar way, the maximum distance between centers of adjacent unit circles is calculated as $\delta_{CP}^N = 1/2R^N$.

Table 1 presents the ρ and δ obtained for learned spaces trained with Contrastive Loss (CL), Triplet Loss (TL), and our proposed implementations (SM-CL and SM-TL). Note that for all N, the proposed SM-CL and SM-TL outperform the previous implementations in both ρ and δ. Larger values of δ mean larger distance between classes (i.e., high inter-class distance), while lower values of ρ mean lower distance between samples from the same class (i.e., low intra-class distance). Table 1 also includes the theoretical maximum distance obtained by the optimal Circle Packing solution. The results show that the proposed distances (SM-CL and SM-TL) outperform the Circle Packing optimal solution δ_{CP}^N with distances close to the theoretical maximum $\delta_{\max}^N$. This is possible as the embeddings of the proposed learned feature spaces tend to be clustered in the border of the unit circle instead of the center.

4. Experiments

The SetMargin Loss has been developed mainly to improve the learned space in open-set classification scenarios (anticipating that it will be also very helpful in many other machine learning scenarios). To evaluate our loss function, we therefore propose to experiment on a challenging *keystroke biometric identification* task, where subjects are identified based on their typing behavior [3,7,34–36].

4.1. Dataset

Our experiments are conducted with the Aalto University Dataset [37] that comprises keystroke sequences from 168,000 subjects. Specifically, we employ the first 78,000 subjects available in the database. The acquisition task asked subjects to memorize English sentences and then to type them as quickly and accurately as they could. The English sentences were selected randomly from a set of 1,525 examples taken from the Enron mobile email and Gigaword Newswire corpus. The example sentences contained a minimum of 3 words and a maximum of 70 characters. Note that the sentences typed by the subjects could contain a bit more than 70 characters because each subject could forget or add new characters when typing. All subjects in the database completed 15 sessions with a different sentence in each session on either a desktop or laptop keyboard. See [37] for more details including demographic and acquisition information.

Table 1
Results of the minimum distance between centroids δ and intra-cluster dispersion ρ (in parenthesis, multiplied by 100) for the different loss functions. We also include the values of δ for the Circle Packing optimal solution (δ_{CP}) and the theoretical maximum ($\delta_{\max}$).

Method	N=12	N=14	N=16	N=18	N=20
δ_{CP}	0.25	0.23	0.22	0.21	0.19
$\delta_{\max}$	0.33	0.30	0.28	0.26	0.24
Contrastive [9]	0.23(0.55)	0.15(0.29)	0.10(1.09)	0.09(1.25)	0.08(1.26)
Triplet [10]	0.22(0.27)	0.15(0.30)	0.09(0.34)	0.08(0.39)	0.08(0.39)
SM-CL [ours]	0.26(0.13)	0.25(0.14)	0.25(0.20)	0.21(0.18)	0.20(0.21)
SM-TL [ours]	0.30(0.15)	0.26(0.18)	0.25(0.19)	0.21(0.20)	0.20(0.21)

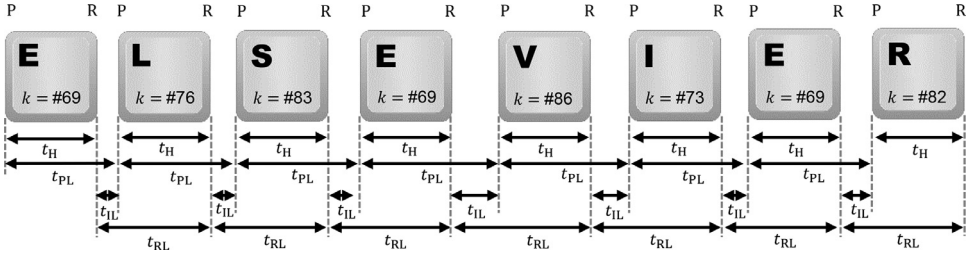

Fig. 5. Example of the 37 temporal features extracted from the keystroke sequence "ELSEVIER": 8 × Hold time (t_H) + 7 × Inter-key Latency (t_{IL}), 7 × Press Latency (t_{PL}), 7 × Release Latency (t_{RL}), 8 × key codes (k). P = key Press event; R = key Release event.

4.2. Pre-processing and keystroke dynamics

The keystroke raw data comprises a three dimensional time series including (see Fig. 5): key press timestamps, key release timestamps, and the keycodes. Timestamps are in UTC format with millisecond resolution, and the keycodes are integers between 0 and 255 according to the ASCII code. The input to the model comprises sequences of keycodes plus 4 temporal features: *(i)* Hold Latency: the elapsed time between press and release key events; *(ii)* Inter-key Latency: the elapsed time between releasing a key and pressing the next key; *(iii)* Release Latency: the elapsed time between two consecutive release events; and *iv)* Press Latency: the elapsed time between two consecutive press events. These 4 features are commonly used in both fixed-text and free-text keystroke biometric systems [38]. This feature extraction process results in a $K \times 5$ feature vector where textcolorBlue$K$ is the number of keys pressed (note that keycode is added for each key pressed). In order to train the model with sequences of different lengths K within a single batch, we truncate the end of the input sequence when $K > M$ and zero pad at the end when $K < M$, in both cases to the fixed size M. The size of the time dimension M was fixed to $M = 50$, which was determined heuristically based on the characteristics of the dataset used in the experiments (see Section 4.1). Error gradients are not computed for the padded zeros which do not contribute to the loss function.

4.3. Implementation details: RNN model and experimental protocol

In our experiments we used the architecture proposed in [7]: TypeNet. TypeNet is a RNN architecture composed of two Long Short-Term Memory (LSTM) layers of 128 units, and an initial Masking layer. Between the LSTM layers, the model performs batch normalization and dropout at a rate of 0.5 to avoid overfitting. Additionally, each LSTM layer has a dropout rate of 0.2. The output of the model $\mathbf{f}(\mathbf{x})$ is an array of size 128 that we use as an embedding feature vector.

In our experiments, a batch was composed of 256 set pairs $(\{\mathbf{x}_k^i\}_{k=1,2,3}, \{\mathbf{x}_q^j\}_{q=1,2,3})$ randomly chosen from the dataset available for learning. The number of possible set pairs is at billions scale. We used 500 batches per epoch. The learning converges with less than 40 epochs which means around 5M set pairs in total.

Training protocol: The RNN model was trained using the first 68,000 subjects in the dataset according to the method proposed in [7]. From the remaining 100,000 subjects, we employed another 10,000 subjects to perform the evaluation of the different loss functions, so there is no data overlap between the two groups of subjects. The distance between two keystroke sequences was computed by averaging the Euclidean distances between the T gallery embedding vectors $\mathbf{f}(\mathbf{x}_g^i)$ and the query embedding vector $\mathbf{f}(\mathbf{x}_q^j)$ as follows:

$$d_{i,j} = \frac{1}{T}\sum_{g=1}^{T} ||\mathbf{f}(\mathbf{x}_g^i) - \mathbf{f}(\mathbf{x}_q^j)|| \tag{9}$$

The experiments include two scenarios: identification and verification. The results reported in the next section are computed in terms of Rank-n and Equal Error Rate (EER). We used the same trained models for both scenarios.

Identification protocol: As a 1:N problem, the identification accuracy varies depending on the size of the background set. The background is conformed with identities (i.e., classes) not used in the learning process (i.e., open-set problem). The goal is to identify the query identity among all the background subjects. In our experiments, the size of the background was equal to 5,000 identities (i.e., subjects). Additionally, our experiments include another 5,000 subjects employed as query set for a total of 10,000 different subjects in testing (5K background + 5K query). We divided the 15 keystroke sequences available for each subject into a gallery set (the first 10 keystroke sequences) and a query set (the remaining 5 keystroke sequences). Remember that it is a free-text scenario, so the model must identify the subject typing different keystroke sequences between gallery and query. We evaluated the identification accuracy by averaging the distance between the query set of samples $\mathbf{x}_q^j$, with $q = 1, \ldots, 5$ belonging to the subject j and the gallery set $\mathbf{x}_g^i$, with $g = 1, \ldots, 10$ belonging to all 5,000 background subjects i. Rank-n is a measure of 1:N identification system performance. A Rank-1 means that $d_{i,J} < d_{I,J}$ for any $i \neq I$. We then identify a query subject J to be present in the gallery as subject I as follows [39]:

$$I = \arg\min_i d_{i,J} \tag{10}$$

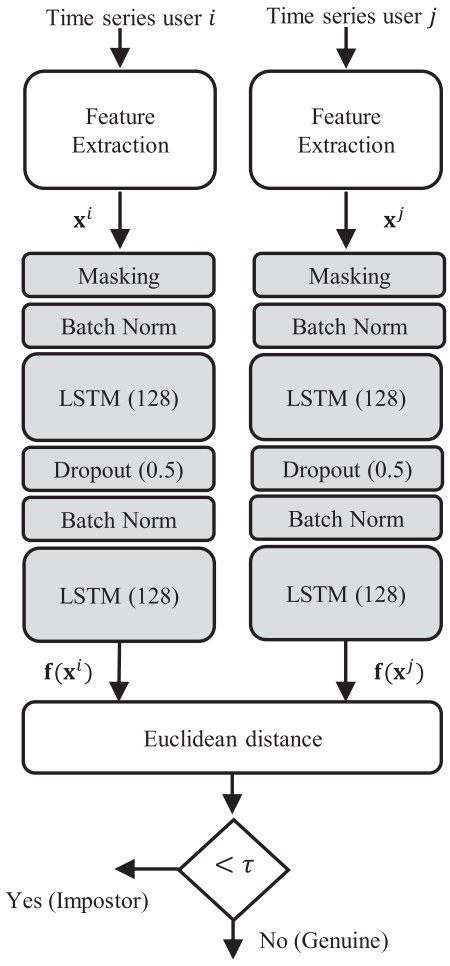

Fig. 6. Block diagram of the keystroke biometric verification system (1 : 1 comparison) based on the TypeNet [7] architecture. τ is a decision threshold.

This identification protocol did not include a decision threshold. The identity with the minimum distance was chosen among all the background subjects.

Verification protocol: The EER is a measure of 1:1 verification system performance. EER is defined as the operational point where False Rejection and False Acceptance are equal. The goal is to verify the identity of a query sample using its corresponding gallery set. This experiment includes a total of 5,000 different subjects in testing. We divided the 15 keystroke sequences available for each subject into a gallery set (the first 5 keystroke sequences) and a query set (the last 5 keystroke sequences). The score $d_{i,j}$ was obtained as the average distance between the query vector and the set of gallery samples. In this experiment, each query sample was evaluated separately for a total number of 50,000 genuine scores (5 query samples $\times$ 5,000 subjects). The impostor scores were obtained choosing one query sample per subject for a total number of 24,995 scores (5,000 $\times$ 4,999). We evaluated the verification accuracy averaging the EER obtained for each subject. Fig. 6 presents the block diagram of the verification protocol for a comparison 1 : 1 $T = 1$.

4.4. Results

Table 2 presents the identification performance of TypeNet [7] incorporating our proposed SetMargin Loss (*SM-CL* and *SM-TL*) in comparison to other popular loss functions: Triplet Loss [10], Contrastive Loss [9], Softmax, Deep Linear Discriminant Analysis DeepLDA [40], Quadruplet Loss [12], and N-Pair Loss [11]. We also include there as reference the results of two competitive algorithms for free-text keystroke biometrics based on statistical models: Partially Observable Hidden Markov Models [6], and Digraphs-SVM [8]. Note that all approaches were trained using the same number of training sequences. Depending on the ML approach, the number of samples employed to compute the loss function varies. In order to make a fair comparison between loss functions, the batch size of the different approaches has been modified to incorporate the same number of samples per batch (1,000 samples per batch).

Table 2
Identification performance in terms of Rank-n accuracy for different methods in the literature. G is the number of samples conforming each set of samples employed to train the SetMargin Loss. Background dataset of 5,000 subjects.

Method	Rank-1	Rank-5	Rank-20
Digraph [8]	0.5%	0.9%	1.2%
POHMM [6]	6.1%	10.3%	13.8%
TypeNet: *Contrastive Loss* [7]	17.8%	31.5%	38.9%
TypeNet: *DeepLDA* [40]	34.2%	63.2%	84.2%
TypeNet: *Softmax*	37.9%	64.9%	84.4%
TypeNet: *Triplet Loss* [10]	38.2%	68.2%	88.5%
TypeNet: *Quadruplet Loss* [12]	38.6%	68.7%	87.9%
TypeNet: *N-Pair Loss* [11]	38.7%	67.7%	87.0%
TypeNet: *SM-CL, G=3*	31.0%	59.9%	82.7%
TypeNet: *SM-CL, G=6*	37.5%	67.0%	86.8%
TypeNet: *SM-CL, G=9*	36.7%	65.8%	86.3%
TypeNet: *SM-TL, G=3*	39.4%	68.3%	88.1%
TypeNet: *SM-TL, G=6*	45.8%	73.9%	91.0%
TypeNet: *SM-TL, G=9*	45.3%	72.4%	89.5%

The results show how Triplet Loss obtains an accuracy two times higher than Contrastive Loss and similar performance than Softmax, Quadruplet Loss [12], and N-Pair Loss [11]. In comparison with traditional statistical approaches [6,8], the Deep Learning model (TypeNet) is clearly superior. The proposed SM-CL obtains a much higher performance than traditional Contrastive Loss but the accuracy achieved is still under other loss functions. Finally, SM-TL improves the best related loss function (N-Pair Loss) by 18% relatively with a Rank-1 accuracy of 45.8%. SM-TL is able to capture the intra-class structure of samples from the same class and at the same time maximizes the inter-class distance. This learning process is appropriate for open-set classification tasks where query samples are matched to multiple different classes not seen during learning (5,000 in our experiments). These accuracies increase up to 73.9% and 91% for Rank-5 and Rank-20 respectively.

The results prove how the performance improves when incorporating sets of samples into the loss function. The superior performance of SM-TL cannot be attributed exclusively to the larger number of samples included in the computation of the loss function. As an example SM-CL showed lower performance with the same number of samples and N-Pair loss showed lower performance with larger number of samples. We have not included the performance of ArcFace in our comparison because of the poor results obtained. This poor performance can be caused by different factors including the low number of samples available per subject (only 15 in contrast with hundreds of samples in [15]), the architecture (RNN instead of CNN), or parameter tuning.

Table 3 presents the verification performance of our proposed SetMargin Loss (*SM-CL* and *SM-TL*) and other popular methods and loss functions. The verification scenario is characterized by higher accuracies in comparison with the identification experiments. In this case, the proposed method (SM-TL) is capable of achieving a performance of 1.85% of EER. The method shows superior performance than previous approaches and popular loss functions.

Table 3
Verification performance in terms of Equal Error Rate (EER) for different methods in the literature. G is the number of samples conforming each set of samples employed to train the SetMargin Loss. The best performance is obtained for $G = 6$.

Method	EER
Digraph [8]	43.1%
POHMM [6]	24.7%
TypeNet: *Contrastive Loss* [7]	5.40%
TypeNet: *DeepLDA* [40]	4.21%
TypeNet: *Softmax*	10.8%
TypeNet: *Triplet Loss* [10]	2.20%
TypeNet: *Quadruplet Loss* [12]	2.33%
TypeNet: *N-Pair Loss* [11]	2.51%
TypeNet: *SM-CL, G=6*	2.42%
TypeNet: *SM-TL, G=6*	1.85%

Nonetheless, in the verification scenario the margin of improvement is lower than in the identification experiment. It should be noted that we used the same model for both the identification and verification scenarios. The results demonstrate the high discrimination capacity of the learned spaces for both identification and verification.

4.4.1. Computational load

Metric Learning approaches suffer from data expansion when batches are conformed by pairs or triplets of samples instead of individual samples. This expansion offers some advantages (e.g. data augmentation), but also increases the computational load. The proposed SM-L learning process is defined by pairs of sets, instead of pairs/triplets of samples. Thus, our method exponentially increase the number of possible combinations. However, the convergence of the learning process is relatively fast and affordable for a personal computer with high specifications. All the experiments presented in this work were made with an Intel Core i7-8760H CPU @ 2.2Ghz, 32 GB RAM, NVIDIA GeForce RTX2080. As an example, the time needed to learn the SM-CL and SM-TL models used in our experiments were 7.6 and 8.8 h, respectively.

5. Conclusions

We have presented a new Distance Metric Learning approach called SetMargin Loss (SM-L). Our approach improves intra-class and inter-class structures in learned spaces, which is specially useful (among other machine learning problems) for open-set classification. We have also introduced the Circle Packing problem as a novel way to gain insights into the feature space of learned representations. A feature space that satisfies the Circle Packing problem guarantees a theoretical maximum inter-class distance given compact intra-class distances. Our experiments suggest that SM-L is capable of obtaining a feature space close to the Circle Packing optimal solution.

We have finally applied SM-L to keystroke biometric Identification using the Aalto University Dataset [37]. Our experiments, conducted over a learning set with typing sequences from 68,000 subjects and evaluated over a testing set with 10,000 subjects, demonstrate the superior performance of the proposed approach over other popular loss functions. The proposed approach showed an accuracy (Rank-1 for identification and EER for verification) significantly superior than traditional statistical methods and 18% better (relatively) than Softmax, Triplet, and N-Pair Losses. The proposed SM-TL approach obtained a Rank-1 accuracy of 45.3% and 1.85% of EER. This performance is still far from the most accurate biometrics modalities, but it provides new opportunities in applications in the digital domain (e.g., online authentication, digital forensics). The EER under 2% obtained for the user verification scenario demonstrates the potential of keystroke dynamics in large-scale user authentication applications. For future work we suggest going deeper in the theory behind the proposed methods in order to seek theoretical limits on the achievable performance and to inspire new learning methods. Finally, we also plan to explore the developed methods in other problems beyond the ones explored here both in supervised and unsupervised learning.

Declaration of Competing Interest

None.

Acknowledgements

This work has been supported by projects: PRIMA (MSCA-ITN-2019-860315), TRESPASS-ETN (MSCA-ITN-2019-860813), BIBECA (RTI2018-101248-B-I00 MINECO), edBB (UAM), and Instituto de Ingenieria del Conocimiento (IIC). A. Acien is supported by a FPI fellowship from the Spanish MINECO.

References

[1] M.L. Ali, K. Thakur, C.C. Tappert, M. Qiu, Keystroke biometric user verification using Hidden Markov Model, in: Proc. of IEEE 3rd International Conference on Cyber Security and Cloud Computing, 2016, pp. 204–209.
[2] J. V. Monaco, Robust keystroke biometric anomaly detection, arXiv preprint arXiv:1606.09075(2016).
[3] A. Morales, J. Fierrez, R. Tolosana, J. Ortega-Garcia, J. Galbally, M. Gomez-Barrero, A. Anjos, S. Marcel, Keystroke biometrics ongoing competition, IEEE Access 4 (2016) 7736–7746.
[4] F. Bergadano, D. Gunetti, C. Picardi, User authentication through keystroke dynamics, ACM Trans. Inf. Forensics Secur. 5 (4) (2002) 367–397.
[5] A. Morales, M. Falanga, J. Fierrez, C. Sansone, J. Ortega-Garcia, Keystroke dynamics recognition based on personal data: a comparative experimental evaluation implementing reproducible research, in: Proc. of IEEE International Conference on Biometrics Theory, Applications and Systems, 2015.
[6] J.V. Monaco, C.C. Tappert, The partially observable hidden Markov model and its application to keystroke dynamics, Pattern Recognit. 76 (2018) 449–462.
[7] A. Acien, J.V. Monaco, A. Morales, R. Vera-Rodriguez, J. Fierrez, TypeNet: scaling up keystroke biometrics, in: Proc. of IEEE/IAPR International Joint Conference on Biometrics, 2020.
[8] H. Aeker, S. Upadhyaya, User authentication with keystroke dynamics in long–text data, in: Proc. of the IEEE International Conference on Biometrics Theory, Applications and Systems, 2016.
[9] R. Hadsell, S. Chopra, Y. Lecun, Dimensionality reduction by learning an invariant mapping, in: Proc. Computer Vision and Pattern Recognition Conference, 2006.
[10] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.
[11] K. Sohn, Improved deep metric learning with multi-class n-pair loss objective, in: Advances in Neural Information Processing Systems, 2016, pp. 1857–1865.
[12] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 403–412.
[13] Y. Wen, K. Zhang, Z. Li, A discriminative feature learning approach for deep face recognition, in: Proc. of the European Conference on Computer Vision, 2006.
[14] W. Liu, Y. Wen, Z. Yu, M. Yang, Large-margin softmax loss for convolutional neural networks, in: Proc. of International Conference on Machine Learning, vol. 2, 2016, pp. 507–516.
[15] J. Deng, J. Guo, X. Niannan, S. Zafeiriou, ArcFace: additive angular margin loss for deep face recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.
[16] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat, R. Adams, Scalable Bayesian optimization using deep neural networks, in: Proc. of the International Conference on Machine Learning, 2015, pp. 2171–2180.
[17] Q. Qi, J. Rong, Z. Shenghuo, L. Yuanqing, Fine-grained visual categorization via multi-stage metric learning, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[18] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, DeepFace: closing the gap to human-level performance in face verification, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.
[19] D. Deb, A. Ross, A.K. Jain, K. Prakah-Asante, K.V. Prasad, Actions speak louder than (pass) words: Passive authentication of smartphone users via deep temporal features, in: Proc. of the IEEE International Conference on Biometrics, 2019.

[20] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[21] O. Rippel, P. Dollar, Metric learning with adaptive density discrimination, in: Proc. of International Conference on Learning Representations, 2016.
[22] A.F. Beardon, K. Stephenson, The uniformization theorem for circle packings, Indiana Univ. Math. J. 39 (4) (1990) 1383–1425.
[23] H. Melissen, Densest packing of eleven congruent circles in a circle, Geometriae Dedicata 50 (1994) 15–25.
[24] F. Fodor, The densest packing of 12 congruent circles in a circle, Beitrge zur Algebra und Geometrie 41 (2000) 401–409.
[25] F. Fodor, The densest packing of 13 congruent circles in a circle, Beitrge zur Algebra und Geometrie 44 (2003) 431–440.
[26] M. Hifi, R. M'Hallah, A dynamic adaptive local search algorithm for the circular packing problem, Eur. J. Oper. Res. 183 (3) (2007) 1280–1294.
[27] M. Hifi, R. M'Hallah, Beam search and non-linear programming tools for the circular packing problem, Eur. J. Oper. Res. 1 (4) (2009) 476–503.
[28] D. Ha, D. Eck, A neural representation of sketch drawings, in: Proc. of International Conference on Learning Representations, 2018.
[29] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. of International Conference on Learning Representations, 2015.
[30] K. Stephenson, Circle packing: a mathematical tale, Not. AMS 50 (11) (2003) 1376–1388.
[31] E.G. Birgin, L.H. Bustamante, H.F. Callisaya, J.M. Martínez, Packing circles within ellipses, Int. Trans. Oper. Res. 20 (3) (2013) 365–389.
[32] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[33] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv preprint arXiv:1703.07737 (2017).
[34] F. Monrose, A. Rubin, Authentication via keystroke dynamics, in: Proc. of the ACM conference on Computer and Communications Security, 1997, pp. 48–56.
[35] D. Gunetti, C. Picardi, Keystroke analysis of free text, ACM Trans. Inf. Syst. Secur. 8 (3) (2005) 312–347.
[36] S.P. Banerjee, D.L. Woodard, Biometric authentication and identification using keystroke dynamics: a survey, J. Pattern Recognit. Res. 7 (1) (2012) 116–139.
[37] V. Dhakal, A.M. Feit, P.O. Kristensson, A. Oulasvirta, Observations on typing from 136 million keystrokes, in: Proc. of the Conference on Human Factors in Computing Systems, 2018.
[38] A. Alsultan, K. Warwick, Keystroke dynamics authentication: a survey of free–text, Int. J. Comput. Sci. Issues 10 (2013) 1–10.
[39] A. Morales, A. Acien, J. Fierrez, J.V. Monaco, R. Tolosana, R. Vera-Rodriguez, J. Ortega-Garcia, Keystroke biometrics in response to fake news propagation in a global pandemic, in: Proc. of IEEE International Workshop on Secure Digital Identity Management, 2020.
[40] L. Wu, C. Shen, A. Van Den Hengel, Deep linear discriminant analysis on fisher networks: a hybrid architecture for person re-identification, Pattern Recognit. 65 (2017) 238–250.

Aythami Morales Moreno M.Sc. in Electrical Engineering, and Ph.D from the Universidad de LPGC in 2006 and 2011. Since 2017, he is Associate Professor with the Universidad Autonoma de Madrid. In his work, he combines his interests in machine learning, biometric processing, security, and privacy.

Julian Fierrez received the M.Sc. and Ph.D. degrees in 2001 and 2006, respectively. He is now Associate Professor at UAM. His research interests include signal and image processing, pattern recognition, security, and biometrics. He received the IAPR Young Biometrics Investigator Award 2017 and is Associate Editor of Elseviers INFORMATION FUSION.

Alejandro Acien received the MSc in Electrical Engineering in 2015 from Universidad Autonoma de Madrid. In October 2016, he joined the Bida Lab group, where he is currently collaborating as an assistant researcher pursuing the PhD degree. He is working in Behaviour Biometrics, HCI, Cognitive Biometric Authentication.

Ruben Tolosana received received the M.Sc. in Telecommunication Engineering, and Ph.D. from Universidad Autonoma de Madrid in 2014 and 2019. His research interests are mainly focused on pattern recognition and machine learning, particularly in the areas of face manipulation and fake detection, human-computer interaction and biometrics.

Ignacio Serna B.S. degree in mathematics and B.S. degree in computer science from the Autonomous University of Madrid, Spain, in 2018, and M.S. degree in Artificial Intelligence in 2020. He is currently pursuing a Ph.D. in Computer Science at the BiDA-Lab. His research interests lie in computer vision, pattern recognition and explainable AI, with applications to biometrics.