Self-Learning with Stochastic Triplet Loss

João Ribeiro Pinto
INESC TEC & Universidade do Porto
Porto, Portugal
joao.t.pinto@inesctec.pt

Jaime S. Cardoso
INESC TEC & Universidade do Porto
Porto, Portugal
jaime.cardoso@inesctec.pt

***Abstract*—Deep learning has offered significant performance improvements on several pattern recognition problems. However, the well-known need for large amounts of labeled data limits applicability and performance where those are not available. Hence, this paper proposes an adaptation of the triplet loss for self-learning with entirely unlabeled data, where there is uncertainty in the generated triplets. The methodology was applied to off-the-person electrocardiogram-based biometric authentication and unconstrained face identity verification tasks, including stress experiments designed to simulate more difficult circumstances. Despite the uncertainty related to the use of unlabeled data, the method was mostly capable of avoiding negatively affecting the model's performance. The promising results show the proposed method can be a viable alternative to supervised learning in cases where only unlabeled data are available. The method is especially suitable for training with continuous stream-based datasets such as on person re-identification in video streams and continuous electrocardiogram-based biometrics.**

I. Introduction

In recent years, deep learning algorithms have offered improved performance over handcrafted methodologies in several pattern recognition tasks. These commonly take advantage of convolutional layers, which enable the autonomous learning of the most relevant features for the task at hand, and use fully-connected layers for more intricate decision boundaries [1]. However, these improvements come with an important drawback: the need for labeled data.

Most tasks where such performance breakthroughs have been achieved are those where researchers have plenty of labeled data at their disposal. The ImageNet dataset enabled the training of deeper models for better performance in the detection and recognition of objects. Similarly, the VGGFace [2] and VGGFace2 [3] datasets helped on the development of improved models for biometric recognition based on face images.

However, for some pattern recognition problems, supervised data is scarce. In most of these cases, even though available data is plenty, the annotation process is cumbersome and/or expensive. This is very frequent in automatic medical image diagnosis tasks, where several imaging exams are usually available but lack specific annotations which typically would need to be offered by experts.

This work was partially financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project "POCI-01-0145-FEDER-028857", and within the PhD grant "SFRH/BD/137720/2018".

Another exemplary application is video surveillance. Given the current ubiquity of surveillance cameras, the availability of data is not a problem. However, the annotation of individuals on the recordings is a long and expensive endeavor. This limits the performance one can attain in these tasks since deeper models will be harder to train.

Yet another key application is continuous biometrics with electrocardiogram (ECG) signals [4]. Deep learning has offered improved performance through increased robustness to signal noise and variability [5]. However, in scenarios with off-the-person signals, the performance still fails to match the use of cleaner on-the-person signals [6]–[8], especially because current off-the-person signal datasets are too few and too small to train larger and deeper models.

Self-learning (SL) has emerged as a promising approach to learn from unlabeled data. Contrarily to unsupervised learning, SL uses contextual or prior information to automatically define labels and tasks, which then support the learning. Generally, self-learning methods are focused on specific applications, including visual representation learning [9], [10], action classification [11], or human motion capture [12], thus including details that restrict their use to those specific tasks.

On the other hand, several self-learning methods use simple low-level tasks for the training, such as ranking samples [13] or recovering masked parts of an image [14], that do not necessarily guarantee that the learned parameters can be useful for high-level tasks. For example, features that prove adequate for the approximate reconstruction of biometric samples may not be good enough to discriminate their identities. Hence, there is currently the need for a more general and capable self-learning methodology.

In this paper, we propose a novel formulation of the triplet loss [15], applicable for self-supervised learning with unlabeled data, unrestricted to specific problems. The triplet loss is particularly suited for this task since it does not require absolute labels for the samples. As samples are combined into triplets, only their relative label information is required. The proposed formulation finds its key application in sequential data scenarios, where mild assumptions about the data acquisition process enable the adoption of a stochastic adaptation of the triplet loss to trigger and sustain the learning.

In the experimental work, the proposed methodology is successfully applied to off-the-person ECG-based biometric authentication tasks, using signals from the University of Toronto ECG Database (UofTDB) [16], and to unconstrained face

identity verification, using the YouTube Faces dataset [17]. Specific stress experiments were conducted on the ECG-based authentication task to evaluate the behavior of the proposed methodology in strain conditions.

Besides this introduction, this paper includes a description of the proposed stochastic triplet loss methodology, in section II, and its applications to ECG and face biometrics, in section III. The experimental details are presented in section IV, while the results and their discussion are presented in section V. The conclusions drawn from this work are presented in section VI.

II. Stochastic Triplet Loss Methodology

The triplet loss [15] uses triplets of data samples to train a network to accurately assess if two samples belong to the same class. Each triplet is composed by an anchor x_a with identity i_a and two other samples, one positive (x_p) and one negative (x_n), where $i_p = i_a \neq i_n$. The three samples are processed, in parallel, by the same network, which returns a learned representation of each of them (y_a, y_p, and y_n). A measure of distance d is then used to compare the anchor-positive ($d_+ = d(y_a, y_p)$) and anchor-negative ($d_- = d(y_a, y_n)$) pairs of representations, which are used in the computation of the triplet loss. The triplet loss for a single triplet can be defined as:

$$\mathcal{L}(x_a, x_p, x_n) = \max(0, \alpha + d_+ - d_-). \tag{1}$$

During training, the goal to decrease the triplet loss will lead the model to adjust its weights to obtain a final representation which brings samples of the same class closer together (reducing d_+), and samples of different classes further apart (increasing d_-). Here, the margin parameter α will contribute to enforce a minimum distance margin between the samples of different classes.

In the standard triplet loss, it is certain that x_p is sampled from the same class as x_a and x_n is sampled from a class different from the class of x_a. This can be rewritten as $P(\mathbb{I}_{i_a}(i_p) = 1) = 1$ and $P(\mathbb{I}_{i_a}(i_p) = 0) = 1$, where $\mathbb{I}_A(x)$ is the indicator function.

With unlabeled samples, there is an uncertainty associated with the generation of the triplets, arising from the possibility of errors during the selection of the positive and negative samples. We generalize the previous assumption by modeling $\mathbb{I}_{i_a}(i_p)$ as a random variable following a Bernoulli distribution with parameter β, *i. e.*, $P(\mathbb{I}_{i_a}(i_p) = 1) = \beta$. Similarly, $\mathbb{I}_{i_a}(i_n)$ is assumed to follow a Bernoulli distribution with parameter γ, *i. e.*, $P(\mathbb{I}_{i_a}(i_n) = 0) = \gamma$.

Assuming the independence of $\mathbb{I}_{i_a}(i_p)$ and $\mathbb{I}_{i_a}(i_n)$, and conditioned on the true identity of the observations in the triplet, the triplet loss follows a multinoulli distribution, with:

$$\mathcal{L}(x_a, x_p, x_n) = \begin{cases} \max(0, \alpha + d_+ - d_-), & \text{with probability } \beta\gamma \\ \max(0, \alpha + d_- - d_-), & \text{with probability } (1-\beta)\gamma \\ \max(0, \alpha + d_+ - d_+), & \text{with probability } \beta(1-\gamma) \\ \max(0, \alpha + d_- - d_+), & \text{with probability } (1-\beta)(1-\gamma) \end{cases} \tag{2}$$

On average, the middle terms in (2) do not contribute to the learning. The last term in the equation negatively impacts the learning. In practice, one would need more data/time to learn under the noisy sampling of the triplets. The parameters β and γ guide the training of the model through the triplet loss, and their values depend on the specificities of the task and the data. In ideal conditions, these should be as close as possible to 1 to approximate the original triplet loss in supervised settings. Lower values would work against the purpose of the triplet loss and diminish its training effectiveness (or, equivalently, increase the difficulty of the training).

The proposed self-learning methodology can be used to train models with unsupervised data. During triplet generation, after the selection of an anchor sample, one can randomly draw one sample from the dataset to serve as negative sample. Assuming a balanced dataset with C classes will give $\gamma = 1 - 1/C$. If C is large, the probability of errors in negative sample selection $p(i_a = i_n)$ will be very low (*e. g.*, 0.1 for $C = 10$ or 0.01 for $C = 100$), and so will be their impact on the training process. More importantly, in practice, prior knowledge allows us to adopt a sampling strategy with a much higher probability of success.

The positive sample can be obtained through the transformation of the anchor according to $x_p = f(x_a)$. The transformation f should be carefully defined in order to change the anchor according to an expected range of intraclass variability but without degrading the underlying label information carried by the sample. The probability β will depend on the degree to which f complies with this need. Similarly to the negative sample selection, prior knowledge of the data may be useful to maximize the probability of success. For example, when dealing with sequential data, choosing a positive sample closer in time to the anchor will increase the probability of both samples sharing the same label. However, the anchor and the positive sample will likely be more similar, which will restrict the model's robustness to intraclass variability. Hence, one should find a trade-off between ensuring intraclass variability and maximizing the probability of success in positive sample generation.

An approximation of the expected value of the loss in (2) can be computed under some simplified conditions. Assuming a setting with two classes C_1 and C_2, with a probability density functions $p_1(x)$ and $p_2(x)$, respectively. If x_a is sampled from either of the distributions, it results in $p_a(x_a) = \pi p_1(x_a) + (1 - \pi)p_2(x_a)$, with $0 \leq \pi \leq 1$.

Setting $x = [x'_a \ x'_p \ x'_n]'$ with a probability density function $p(x) = p(x_a, x_p, x_n)$ assumed to be equal to

$$\begin{aligned} p(x) &= p(x_a, x_p, x_n) \\ &= p_a(x_a)p_p(x_p|x_a)p_n(x_n|x_a) \\ &= \pi p_1(x_a)p_1(x_p)p_2(x_n) \\ &\quad + (1-\pi)p_2(x_a)p_2(x_p)p_1(x_n) \end{aligned} . \tag{3}$$

The triplet loss between x_a, x_p, x_n can be described using a Euclidean distance function as $\mathbb{E}_{x\sim p}\mathcal{L}(x_a, x_p, x_n)$ with

$$\begin{aligned} \mathcal{L}(x_a, x_p, x_n) = \ &\max\big(0, \alpha + ||r(x_a) - r(x_p)||^2 \\ &-||r(x_a) - r(x_n)||^2\big) \end{aligned} , \tag{4}$$

where $r(x)$ is the learned representation of x.

In the presence of the assumed noise model in the sampling process of the triplets (x_a, x_b, x_c), the probability density function becomes

$$\begin{aligned} g(x) = \ &\beta\gamma p_a(x_a)p_p(x_p)p_n(x_n) \\ &+ (1-\beta)\gamma p_a(x_a)p_n(x_p)p_n(x_n) \\ &+ \beta(1-\gamma)p_a(x_a)p_p(x_p)p_p(x_n) \\ &+ (1-\beta)(1-\gamma)p_a(x_a)p_n(x_p)p_p(x_n) \end{aligned} . \tag{5}$$

With this, the triplet loss becomes

$$\begin{aligned} \mathbb{E}_{x\sim g}\mathcal{L}(x_a, x_p, x_n) = \ &\beta\gamma\mathbb{E}_{x\sim p}\mathcal{L}(x_a, x_p, x_n) \\ &+ (1-\beta)\gamma\mathbb{E}_{x\sim h_1}\mathcal{L}(x_a, x_p, x_n) \\ &+ \beta(1-\gamma)\mathbb{E}_{x\sim h_2}\mathcal{L}(x_a, x_p, x_n) \\ &+ (1-\beta)(1-\gamma)\mathbb{E}_{x\sim p}\mathcal{L}(x_a, x_n, x_p) \end{aligned} , \tag{6}$$

with

$$\begin{aligned} h_1(x_a, x_p, x_n) = \ &\pi p_1(x_a)p_2(x_p)p_2(x_n) \\ &+ (1-\pi)p_2(x_a)p_1(x_p)p_1(x_n) \end{aligned} \tag{7}$$

and

$$\begin{aligned} h_2(x_a, x_p, x_n) = \ &\pi p_1(x_a)p_1(x_p)p_1(x_n) \\ &+ (1-\pi)p_2(x_a)p_2(x_p)p_2(x_n) \end{aligned} . \tag{8}$$

Noting that the expected value of the *gradient* of the loss $\mathcal{L}(x_a, x_p, x_n)$ is zero under h_1 and h_2 (since x_p and x_n are sampled from the same distribution and the loss is symmetric), the impact of those two cases in a gradient based learning scheme is small.

The total loss is then:

$$\begin{aligned} &\beta\gamma \max(0, \alpha + ||y_a - y_p||^2 - ||y_a - y_n||^2) + \\ &(1-\beta)(1-\gamma) \max(0, \alpha + ||y_a - y_n||^2 - ||y_a - y_p||^2) \end{aligned} , \tag{9}$$

under the $p(x)$ probability density function, where $y = r(x)$.

Finally, this loss can be compacted to:

$$\begin{aligned} &\max\big(0, \beta\gamma(\alpha + ||y_a - y_p||^2 - ||y_a - y_n||^2)\big) + \\ &\max\big(0, (1-\beta)(1-\gamma)(\alpha + ||y_a - y_n||^2 - ||y_a - y_p||^2)\big) \end{aligned} . \tag{10}$$

In section III, example applications of this methodology are presented for the tasks of electrocardiogram-based biometric authentication and face identity verification.

III. Applications

The proposed method can be used to train models relying solely on unsupervised data. On classification tasks, the negative sample can be generated through the random selection of a sample in the dataset. Assuming a large number of balanced classes, errors in negative sample selection should be rare. For the positive samples, the function $f(x)$ that generates them based on an anchor can be a data augmentation procedure. This should be carefully adjusted to cover the expected intraclass noise and variability, while retaining the information pertaining to the underlying image label, which can be difficult.

Alternatively, when training with sequential data, the triplet generation can forgo the data augmentation procedures. In these situations, depending on the acquisition context or protocol, the temporal distance or proximity between data can be used to infer the identity of the subjects. A sample which is very close in time to the anchor can safely be used as x_p. Similarly, a sample which is sufficiently distant in time to the anchor can be assumed to belong to a different user, and thus used as x_n. Some knowledge of the domain and the acquisition settings can be used to adjust the distance between x_a, x_p, and x_n to maximize β and γ.

Both aforementioned alternatives (entirely unsupervised or using sequential data) were explored for the applications described below, through the experiments described in section IV.

A. ECG-based Biometric Authentication

Deep learning models have previously shown improved robustness to off-the-person noise and variability in electrocardiogram-based biometrics [5], [7]. However, to train such models and match the performances reported for cleaner on-the-person signals, one would need large databases of off-the-person acquisitions, which are currently unavailable [4]. In such circumstances, a pretrained network would often be the natural option in computer vision tasks. However, these too are currently nonexistent for unidimensional physiological signals as the electrocardiogram.

The integration of ECG sensors in everyday objects, *e. g.* using the CardioWheel steering wheel cover [18] for shared vehicles or similar solutions for shared bicycles or scooters, enables the continuous acquisition of data from several subjects over long periods. This large amount of collected data could be used to train deeper and more sophisticated models. However, this data is commonly unlabeled, as the identity of the users at the moment of acquisition cannot be easily verified.

The proposed methodology for self-learning can be applied to train models for ECG-based authentication using such data. As aforementioned, perturbations based on data augmentation procedures can be applied to the anchor to generate a positive sample. Thus, the four most successful data augmentation procedures proposed by Pinto *et al*. [5] were implemented. For each triplet, one of these was randomly selected to generate a positive sample from the anchor:

- *Cropping*: a smaller contiguous segment is taken from the anchor sample and resampled to match the anchor's length, to simulate slower heart rates;
- *Baseline Wander*: a periodic undulation, with a frequency near 1 Hz, is added to the anchor segment to simulate breathing movement artifacts;
- *Gaussian Noise*: Gaussian noise is added to the anchor signal, simulating high-frequency distortions similar to the electromyogram (EMG) and powerline interference;
- *Random Permutation*: the anchor is divided into N subsegments, which are shuffled to generate a different sample that simulates discontinuities or sensor faults.

With continuous ECG recordings, it is possible to avoid errors in positive and negative sample selection. Having separate recordings for each person, positive samples are obtained through the selection of a segment of the anchor's recording. The negative sample is obtained from a different recording. In this case, there should be no errors in positive sample selection. Although there can be several recordings for the same person, errors in negative sample selection should be rare considering the large number of identities in the dataset and the balanced number of recordings per identity.

B. Face Identity Verification

More face data are available now than ever before, especially from surveillance feeds or public videos shared in online social media platforms. However, as with ECG-based biometrics, the labeling of faces in acquired datasets is a tedious and lengthy task. Some researchers have taken advantage of online videos to build large datasets for face recognition, such as the YouTube Faces dataset from Wolf *et al*. [17]. However, these datasets are limited by the amount of annotations available.

The proposed self-learning method can be used to train models for face verification without labeled data. In this case, common image data augmentation based on rotations, width and height shifts, and horizontal flips were used as the transformation function $f(x)$ that generates a positive sample x_p based on an anchor x_a.

Having short videos, a random detected face from the same recording as the anchor can serve as a positive sample, while a negative sample can be drawn from a different recording. With some knowledge of the recordings, we minimize the probability of errors in positive and negative sample selection. Specifically, we know the YouTube Faces data consists of frames from short video recordings, with several people, with no more than one person per frame. Hence, although the short recordings lack much intrasubject trait variability, selecting triplets in the aforementioned way avoids errors in positive and negative sample selection.

IV. Experimental Settings

A. Data

1) ECG data: The data used to train and evaluate the model is from the University of Toronto ECG Database (UofTDB) [16]. This database includes data from 1019 subjects, acquired at 200 Hz using dry metallic button electrodes, held by the subjects in contact with one finger of each hand. Each recording is $2-5$ minutes long, and each subject has recordings for up to five different postures (supine, tripod, exercise, standing, and sitting) on up to six sessions over a period of six months.

The data was divided for model training and evaluation as done by Pinto *et al*. [7]. The last 100 subjects (from subject 921 to subject 1020) were reserved for model training. The data from the remaining 918 subjects were used for evaluation. One subject (8) was discarded for having too few data. From the 918 subjects reserved for evaluation, the first 30 seconds of the first recording were used for enrollment, while the remaining data were used for testing. This aimed to mimic a realistic context with scarce supervised data as expected in real ECG-based biometric applications.

2) Face data: For face identity verification, data from the YouTube Faces database [17] were used. This database contains frames from 3425 videos of 1595 subjects, sourced from YouTube. Each video is 48 to 6070 frames long, and there are up to six videos of each subject. This work used the aligned images provided on the database, which resulted from face detection, cropping, and alignment.

The first 150 subjects (in alphabetical order) were used to build the dataset used in this work: the first 100 subjects were reserved for training and validation, while the data from the remaining subjects were used for testing. Triplets were generated using this data subset, after resizing the images to 224×224, as detailed below in the experiments' description.

B. Models

The self-supervised training method proposed in this paper was explored for ECG-based authentication using an adapted version of the end-to-end network proposed by Pinto *et al*. [7] (see Fig. 1). The model receives two z-score normalized five-second raw ECG segments (a stored template and a query sample) and returns a measure of dissimilarity related to their identity.

The network is composed of four convolutional layers followed by a dense layer. A max-pooling layer (pooling size 1×5) follows each of the first three convolutional layers. The convolutional layers have 16, 16, 32, and 32 filters, respectively, with unit stride, without padding. The dense layer is composed of 100 units. All convolutional and dense layers are followed by ReLU activation.

For face identity verification, the model is a simple convolutional neural network (see Fig. 2), which receives two 224×224 RGB face images, normalized to $[0, 1]$ intensities, and outputs a measure of their dissimilarity. It is composed of six convolutional layers interposed with five max-pooling layers (pooling size 2×2) and followed by two dense layers. The convolutional layers have 16, 16, 32, 32, 64, and 64 filters, respectively, with size 3×3, unit stride, without padding. The dense layers are composed of 1000 and 100 units, respectively. All convolutional and dense layers are followed by ReLU activation.

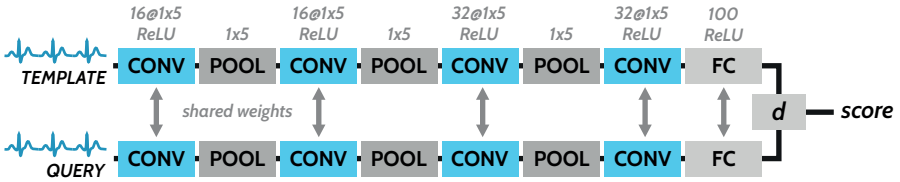

Fig. 1. Architecture of the ECG authentication model, adapted from [7], that was trained with the proposed methodology.

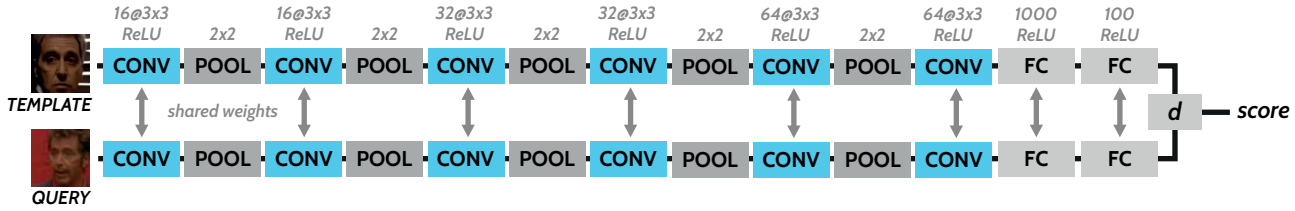

Fig. 2. Architecture of the face identity verification model that was trained with the proposed methodology.

Both models were trained using the Adam optimizer, with initial learning rate 0.0001. As in [7], the Euclidean distance was used as distance measure d during training, while for authentication this was replaced by the normalized Euclidean distance for scores in $[0, 1]$. The triplet loss margin was set as $\alpha = 1.0$. A maximum of 200 epochs was given, with batches of 12 triplets, along with early stopping with patience of 10 epochs. Dropout was used before each dense layer, with rates 0.5 and 0.2 for the ECG and the face models, respectively. L2 regularization ($\lambda = 0.01$) was used for the convolutional layers in both models.

C. Evaluation Metrics

The evaluation metrics used are the False Acceptance Rate (FAR), the False Rejection Rate (FRR), the Equal Error Rate (EER) and the Receiver-Operating Characteristic (ROC) curve [4]. The FAR measures the rate at which impostors meet a given acceptance threshold and are falsely granted access. The FRR measures the rate at which genuine users are incorrectly denied access due to their scores not meeting the given threshold. The EER corresponds to the error at the operation threshold where FAR and FRR have equal values. The ROC curve plots the values of $1-$FRR *versus* FAR for the possible range of threshold values.

D. Experiments' Description

1) Without Supervision: In this experiment, the models were trained with triplets whose negative samples are drawn randomly from the entire respective dataset. The positive samples are created through the application of data augmentation procedure to the respective anchor samples. To train the ECG authentication model, $100\ 000$ triplets were generated for the training, of which 10% were used for validation during training, and $10\ 000$ triplets were generated for evaluation. For the face model, $10\ 000$ triplets were generated for the training, of which 20% were used for validation, and 5000 triplets were generated for testing.

Naturally, depending on the dataset used for training, the probability of error in the random selection of a negative sample will vary. In datasets with fewer classes, the probability of randomly selecting a negative sample whose class matches that of the anchor is greater than in datasets with more classes. Hence, the aforedescribed experiment with the ECG authentication model was repeated, but giving the selection of a negative sample a probability $p_e = 1 - \gamma$ of returning a sample from the anchor's identity. This probability of error was linked to a simulated number of subjects N_s, with $p_e = 1/N_s$ and $N_s = \{2, 5, 10, 20, 50, 100, 200, 500, 1000\}$. This enabled the assessment of how a balanced dataset with fewer classes could impact the training process and the effect on the final authentication performance.

2) Using Recordings: This experiment used the recordings of the UofTDB database and the video recordings of YouTube Faces as a way to infer the identity of the samples through the temporal proximity between them, using prior knowledge to minimize triplet generation errors. Here, the positive sample is drawn from the same recording as the anchor, while the negative sample is drawn from a different recording. As each subject can have several recordings, there is an error associated with the selection of the negative sample, which can accidentally be selected from a different recording of the same subject. The number of generated ECG and face triplets used for training, validation, and testing, was the same as aforementioned in IV-D1.

When training the network with longer recordings spanning several users, as described in section II, the possible errors are different. Although the positive sample is selected in the temporal vicinity of the anchor, it can belong to a different identity. The negative sample, despite the distance from the anchor, can accidentally belong to the same user as the anchor. Hence, additional experiments were conducted where the positive and negative sample selection processes failed purposely with probability $p = \{0.05, 0.1, 0.2, 0.3, 0.5, 0.7\}$, to assess the effect of such errors in the final model performance.

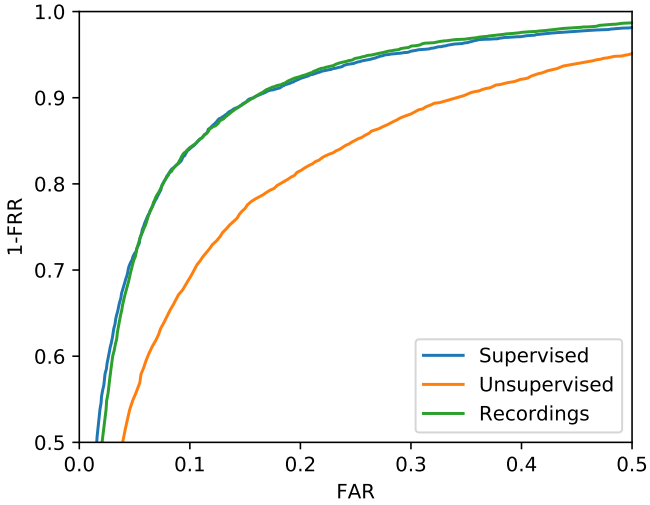

Fig. 3. Comparison of the Receiver-Operating Characteristic curves on ECG authentication for supervised training, unsupervised training, and recording-based supervision (FAR: False Acceptance Rate; FRR: False Rejection Rate).

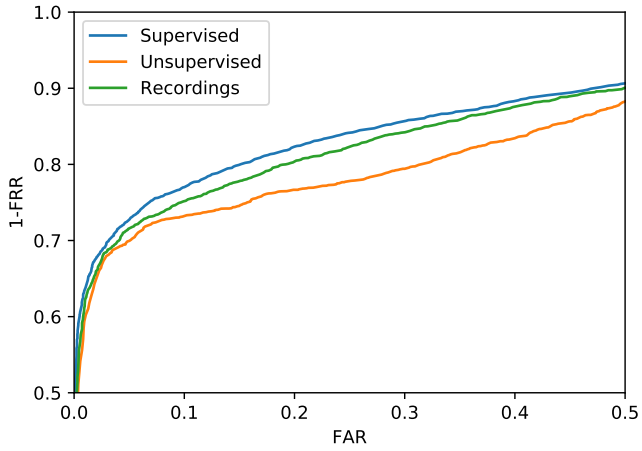

Fig. 4. Comparison of the Receiver-Operating Characteristic curves on face verification for supervised training, unsupervised training, and recording-based supervision (FAR: False Acceptance Rate; FRR: False Rejection Rate).

V. Results and Discussion

A. ECG-based authentication

The baseline results correspond to the authentication model trained with supervised data. The equal error rate of 12.56% represents a small improvement over the corresponding result reported in [7]. Considering the evaluation data and conditions were the same, this method also offered significantly better results than the state-of-the-art methods implemented and tested in [7]: the Autoencoder-based solution proposed by Eduardo *et al*. [19], the AC/LDA method proposed by Agrafioti *et al*. [20], and the DCT approach proposed by Pinto *et al*. [5], [21].

The performance results of the model trained with the two unsupervised training approaches are presented in Fig. 3, in comparison with the baseline results. The Equal Error Rate values were 12.56%, 19.19%, and 12.70%, for supervised, entirely unsupervised, and recording-based training, respectively. The difference between the performance with entirely unsupervised training and the performance with recording-based training denotes the data augmentation procedures have not been able to completely mimic the variability of the signals, and could perhaps be improved using optimized data augmentation [22], [23]. Despite the worse performance attained with the entirely unsupervised training approach, all of these methods offered better performance than the handcrafted methods evaluated in the same settings in [7], among which the best result was 21.82% EER.

B. Face Identity Verification

As with the ECG-based authentication task, the model trained with supervised data was used as a baseline for comparison of results in face identity verification. The performance offered by the baseline was 18.45% EER. This is considerably higher than the state-of-the-art, which is explained by the relative simplicity of the implemented model and the relatively small dataset used. Nevertheless, the goal of this work was not to overcome or match the state-of-the-art in face verification, but to illustrate how the proposed self-learning methodology can be applied to face biometrics with small performance losses relative to a supervised baseline in similar conditions.

The results with the proposed methodology are illustrated in Fig. 4. When using entirely unsupervised data, the proposed method offered 22.81% EER, a 4.36% increase relative to the use of supervised data. With recording-based triplet generation, the model offered 19.77% EER, a 1.32% increase. These small performance losses when forgoing labels during training show that the model can learn without supervision using the stochastic triplet loss, as verified above for ECG biometrics. Besides these two applications, one should expect the proposed self-learning methodology to be successfully applicable to similar problems.

C. Stress Experiments

As discussed in subsection IV-D, the success of the unsupervised triplet generation technique depends on the number of identities (classes) on the database. Hence, an experiment was performed on ECG authentication to simulate the variation of the number of identities on the dataset, inducing errors in the negative sample selection with the respective probability. The results (see Fig. 5) show that, although the performance worsens with fewer subjects, the errors have a very small effect for datasets with more than 20 subjects. In fact, with 50 subject or more, the performance results stabilized around 20% EER. Hence, a dataset with 50 classes should be enough to adequately apply this method with better performance than handcrafted state-of-the-art approaches.

For the training based on temporal proximity between samples, both the selection of positive samples and the selection of negative samples may fail. Hence, enforcing a probability of each error in the recording-based training experiments allows

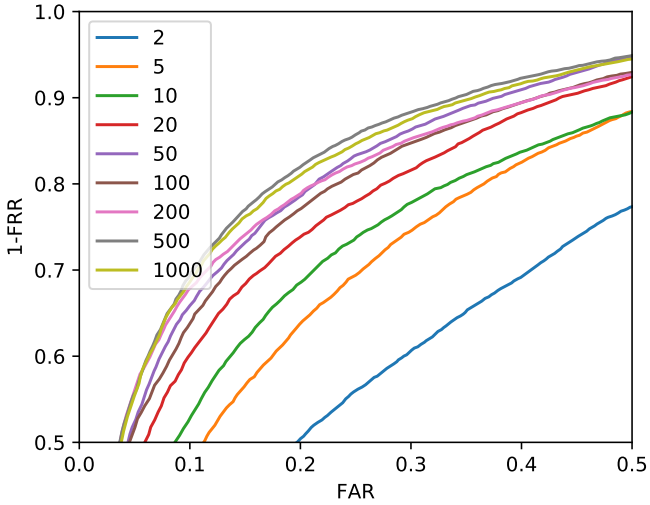

Fig. 5. Receiver-Operating Characteristic curve for negative selection error based on number of database subjects (FAR: False Acceptance Rate; FRR: False Rejection Rate).

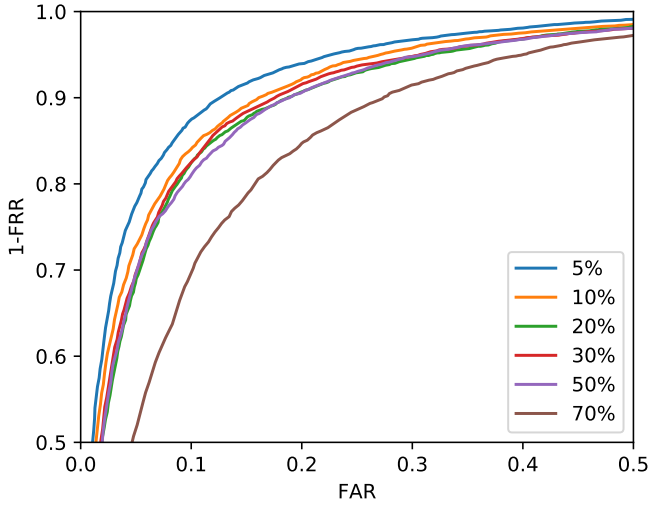

Fig. 7. Receiver-Operating Characteristic curves for varying negative sample selection error probability (FAR: False Acceptance Rate; FRR: False Rejection Rate).

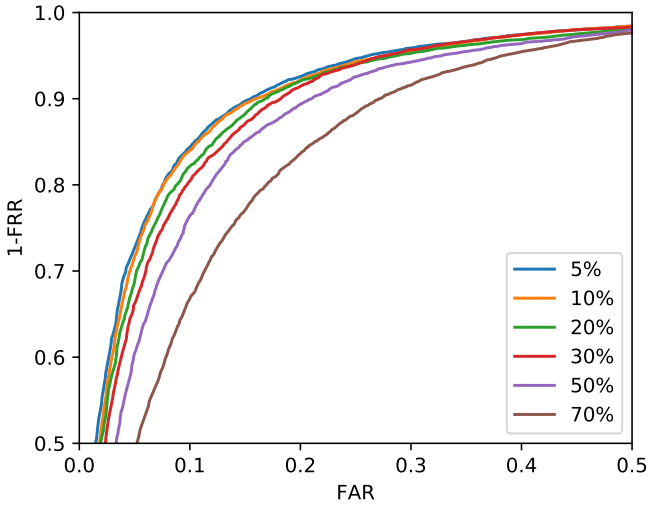

Fig. 6. Receiver-Operating Characteristic curves for varying positive sample selection error probability (FAR: False Acceptance Rate; FRR: False Rejection Rate).

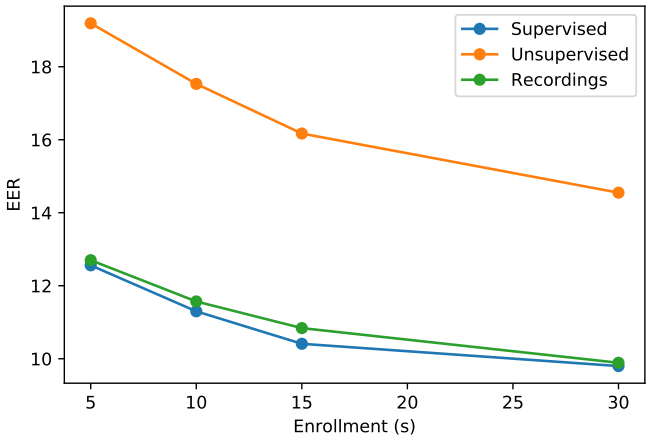

Fig. 8. Equal Error Rate (EER) results when using more enrollment data from each subject.

the study of the impact of such errors in the model's performance. The increase of either positive or negative sample selection error probabilities lead to a decrease in performance (see Fig. 6 and Fig. 7). However, that decrease is small unless the probabilities of error are over 50%. This means that some knowledge of the typical usage times and patterns during data acquisition would be enough to adjust the process of positive and negative sample selection and ensure the best results.

Results can be further improved using more enrollment data (see Fig. 8). As studied by Pinto *et al*. [7], instead of the simple one-vs-one comparisons performed in the aforedescribed experiments, which correspond to five-second enrollments, the query template can be compared with each of several enrollment templates from each person, and only the minimum score is considered. ECG authentication performance with the proposed unsupervised and recording-based training approaches reached 14.55% and 9.89%, respectively, when using thirty-second enrollments.

VI. Conclusion

In this paper, a novel formulation of the triplet loss is proposed for self-supervised learning with unlabeled data. This method considers the uncertainty associated with the triplet generation in unsupervised settings, and maximizes probability of success using prior knowledge.

The proposed methodology was applied to the task of ECG-based biometric authentication, using transformations based on data augmentation or the temporal proximity between samples to generate valid triplets. The method offered better

performance than handcrafted state-of-the-art methods, especially when using temporal proximity between samples, with performance results similar to supervised training.

This pattern was also confirmed on the task of unconstrained face identity verification. Training with entirely unsupervised data using the proposed triplet loss formulation resulted in just a small performance loss when compared with the use of supervised data. When generating triplets based on video streams, this loss was considerably smaller.

Hence, although the proposed method can be influenced by errors in the unsupervised triplet generation, its robustness avoided impact to performance in most cases. Thus, this method would, according to the presented results, be a viable training option in multiclass classification problems where only unlabeled data are available, especially with sequential data.

ACKNOWLEDGMENT

The authors wish to acknowledge the creators and administrators of the UofTDB database (University of Toronto, Canada) and the YouTube Faces database (Tel Aviv University, Israel), which were essential for this work.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, 2015.

[2] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in *British Machine Vision Conference (BMVC)*, 2015.

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces across Pose and Age," in *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 05 2018, pp. 67–74.

[4] J. R. Pinto, J. S. Cardoso, and A. Lourenço, "Evolution, Current Challenges, and Future Possibilities in ECG Biometrics," *IEEE Access*, vol. 6, pp. 34 746–34 776, 2018.

[5] ——, "Deep Neural Networks For Biometric Identification Based On Non-Intrusive ECG Acquisitions," in *The Biometric Computing: Recognition and Registration*, K. V. Arya and R. S. Bhadoria, Eds. Boca Raton FL, United States: CRC Press, 2019, ch. 11, pp. 217–234.

[6] E. J. da Silva Luz, G. J. P. Moreira, L. S. Oliveira, W. R. Schwartz, and D. Menotti, "Learning Deep Off-the-Person Heart Biometrics Representations," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1258–1270, May 2018.

[7] J. R. Pinto and J. S. Cardoso, "An End-to-End Convolutional Neural Network for ECG-Based Biometric Authentication," in *10th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Tampa, FL, United States, 2019.

[8] Q. Zhang, D. Zhou, and X. Zeng, "HeartID: A Multiresolution Convolutional Neural Network for ECG-Based Biometric Human Identification in Smart Health Applications," *IEEE Access*, vol. 5, pp. 11 805–11 816, 2017.

[9] Z. Feng, C. Xu, and D. Tao, "Self-Supervised Representation Learning by Rotation Feature Decoupling," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] L. Gomez, Y. Patel, M. Rusinol, D. Karatzas, and C. V. Jawahar, "Self-Supervised Learning of Visual Features Through Embedding Images Into Text Topic Spaces," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[11] B. Fernando, H. Bilen, E. Gavves, and S. Gould, "Self-Supervised Video Representation Learning With Odd-One-Out Networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] H.-Y. F. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki, "Self-supervised Learning of Motion Capture," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Dec. 2017.

[13] X. Li, S. Liu, S. De Mello, X. Wang, J. Kautz, and M.-H. Yang, "Joint-task self-supervised learning for temporal correspondence," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., 2019, pp. 317–327.

[14] T. H. Trinh, M. Luong, and Q. V. Le, "Selfie: Self-supervised pretraining for image embedding," *CoRR*, vol. abs/1906.02940, 2019. [Online]. Available: http://arxiv.org/abs/1906.02940

[15] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large Scale Online Learning of Image Similarity Through Ranking," *Journal of Machine Learning Research*, vol. 11, pp. 1109–1135, 2010.

[16] S. Wahabi, S. Pouryayevali, S. Hari, and D. Hatzinakos, "On Evaluating ECG Biometric Systems: Session-Dependence and Body Posture," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 11, pp. 2002–2013, Nov. 2014.

[17] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR 2011*, June 2011, pp. 529–534.

[18] A. Lourenço, A. P. Alves, C. Carreiras, R. P. Duarte, and A. Fred, "CardioWheel: ECG Biometrics on the Steering Wheel," in *Machine Learning and Knowledge Discovery in Databases*, A. Bifet, M. May, B. Zadrozny, R. Gavalda, D. Pedreschi, F. Bonchi, J. Cardoso, and M. Spiliopoulou, Eds. Cham: Springer International Publishing, 2015, pp. 267–270.

[19] A. Eduardo, H. Aidos, and A. L. N. Fred, "ECG-based Biometrics using a Deep Autoencoder for Feature Learning: An Empirical Study on Transferability," in *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, Porto, Portugal, Feb. 2017, pp. 463–470.

[20] F. Agrafioti, F. M. Bui, and D. Hatzinakos, "Secure Telemedicine: Biometrics for Remote and Continuous Patient Verification," *Journal of Computer Networks and Communications*, vol. 2012, p. 11, 2012.

[21] J. R. Pinto, J. S. Cardoso, A. Lourenço, and C. Carreiras, "Towards a Continuous Biometric System Based on ECG Signals Acquired on the Steering Wheel," *Sensors*, vol. 17, no. 10, p. 2228, 2017.

[22] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning Augmentation Strategies From Data," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, "Fast AutoAugment," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.