



Locality preserving binary face representations using auto-encoders

Mohamed Amine Hmani, Dijana Petrovska-Delacrétaz, Bernadette Dorizzi

► To cite this version:

Mohamed Amine Hmani, Dijana Petrovska-Delacrétaz, Bernadette Dorizzi. Locality preserving binary face representations using auto-encoders. IET Biometrics, 2022, 11 (5), pp.445-458. 10.1049/bme2.12096 . hal-03944066

HAL Id: hal-03944066

<https://hal.science/hal-03944066v1>

Submitted on 17 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC0 - Public Domain Dedication 4.0 International License

ORIGINAL RESEARCH

Locality preserving binary face representations using auto-encoders

Mohamed Amine Hmani  | Dijana Petrovska-Delacrétaz | Bernadette Dorizzi

Laboratoire SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, Palaiseau, France

Correspondence

Mohamed Amine Hmani, Laboratoire SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, 9 rue Charles Fourier, Évry-Courcouronnes Cedex, Palaiseau, 91011, France.
Email: Mohamed.hmani@telecom-sudparis.eu

Funding information

European Commission, Grant/Award Numbers: 653586, 769872

Abstract

Crypto-biometric schemes, such as fuzzy commitment, require binary sources. A novel approach to binarising biometric data using Deep Neural Networks applied to facial biometric data is introduced. The binary representations are evaluated on the MOBIO and the Labelled Faces in the Wild databases, where their biometric recognition performance and entropy are measured. The proposed binary embeddings give a state-of-the-art performance on both databases with almost negligible degradation compared to the baseline. The representations' length can be controlled. Using a pretrained convolutional neural network and training the model on a cleaned version of the MS-celeb-1M database, binary representations of length 4096 bits and 3300 bits of entropy are obtained. The extracted representations have high entropy and are long enough to be used in crypto-biometric systems, such as fuzzy commitment. Furthermore, the proposed approach is data-driven and constitutes a locality preserving hashing that can be leveraged for data clustering and similarity searches. As a use case of the binary representations, a cancellable system is created based on the binary embeddings using a shuffling transformation with a randomisation key as a second factor.

1 | INTRODUCTION

The face is one of the most widely used biometric characteristics. With the availability of huge face recognition data sets [1, 2] and growing computational power, face recognition performance keeps improving [3–7]. Face recognition has seen vast adoption thanks to its accuracy and ease of use. From smartphones and computers to CCTV cameras and surveillance, face recognition is present everywhere. This widespread presence raises privacy and security concerns. A solution to these concerns is to employ biometric template protection schemes such as crypto-systems and cancellable biometrics. However, to protect the face templates, most of the techniques employed need a binary representation of the face. In addition, most face verification systems employ continuous representations, which are less suitable for template protection schemes.

The major contribution of this paper is introducing a data-driven template binarisation method using Deep Neural Networks (DNN), which does not degrade the performance of the

baseline system. Furthermore, we seek to obtain long binary representations with high entropy to be used in crypto-biometric key regeneration schemes. The proposed binarisation method has four main advantages:

- The degradation of the recognition performance caused by the binarisation is negligible compared to that of the baseline system.
- The binarisation method can be applied to any type of real representation.
- The length of the binary representation can be controlled. The binarisation method provides arbitrary length representations that are limited only by the quality of the training database (size, noise). This allows for flexible representations that can be adapted to multiple applications, such as crypto-biometric key regeneration, fuzzy commitment, and fuzzy extraction schemes.
- The binarisation method keeps the topology of the original space, which allows for the use of the binary representation in database searches and clustering.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *IET Biometrics* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

The paper is organised as follows: In Section 2, we provide a brief survey of biometric binarisation techniques. Section 3 explains the different approaches we followed to extract binary embeddings directly using DNN. Section 4 describes the databases used in this paper. In Section 5, we analyse the performance of the binary representations in terms of biometric recognition and entropy. In Section 6, we study a use case for the binary embeddings consisting in creating a cancellable biometric system using a shuffling transformation as a protection scheme. Finally, the conclusions are laid out in Section 7.

2 | RELATED WORKS

State-of-art face recognition systems use continuous vector embeddings to represent the users. However, the majority of biometric template protection schemes and crypto-biometric systems need a binary representation [8] as an input. Thus, the continuous vectors need to be binarised. Binarisation methods fall into two categories: data-independent and data-dependent strategies.

For the data-independent approaches, different schemes were proposed. In Ref. [9], Drozdowski et al. benchmark data-independent binarisation methods such as Refs. [10–17]. These rule-based methods directly quantise the projected values with a threshold or use an orthogonal matrix to obtain the binary codes. Such methods do not preserve the locality structure in the whole learning process.

As for data-dependent approaches, recently, multiple binarisation techniques based on neural networks such as Refs. [18–20] were introduced. These techniques focus on projecting the input on a predetermined space. For example, in Ref. [18], the authors map a low-density parity-check (LDPC) code to each identity in the training data set. Thus, each person in the training set has their codeword, resulting in perfect discrimination between the training subjects. Nevertheless, the system's performance degrades when enrolling a user that did not belong to the training set.

Pandey et al. [21] use deep convolutional neural networks to learn mapping from face images to maximum entropy binary codes. The mapping is robust enough to tackle the problem of exact matching, yielding the same code for new samples of a user as the code assigned during training. These codes are then hashed to generate protected face templates.

In Ref. [22], Jindal *et al.* generate unique binary codes with maximum entropy. In order to maximise the entropy of the binary codes, each bit of the binary code is randomly generated and has no correlation with the original biometric sample. The binary codes are used to replace the one-hot encoding used to train the VGG-Face network. The network uses binary cross-entropy as the loss function, with the last layer activation function being the sigmoid function instead of the softmax function.

Similar to our approach, Carreira et al. [23] use auto-encoders for the binarisation of the data. The outputs of the hidden layer are passed into a step function to binarise the codes. Incorporating the step function in the learning leads to a non-smooth objective function. Optimising this non-smooth

function is NP-complete. Where the gradients do exist, they are zero nearly everywhere. They use binary SVMs to learn the model parameters to handle this difficulty. Whereas, in our case, we ignore the gradient of the binarisation layer to keep the non-zero aspect of the gradient of the loss function.

The previously mentioned binarisation methods provide binary representations with limited length. In this paper, we aim to obtain long representations with high entropy to be used in crypto-biometric key regeneration.

As opposed to the methods that use a predefined mapping space, the approach we present aims to preserve the topology of the embeddings provided by the baseline DNN architecture. As a result, we preserve the advantages of the underlying DNN (resistance to noise, higher accuracy, and robustness) while obtaining binary representations. Furthermore, persevering the topology of the data also allows for using our binarisation method in data retrieval applications.

In the following sections, we introduce our binarisation method. Then, we study its performance and present a use case of a cancellable biometric system based on the binary representations created using our method.

3 | PROPOSED FACE BINARISATION METHOD

This study uses deep neural networks to extract binary biometric representations from face images. This way, we take advantage of data-driven approaches to generate an optimised binary representation.

Our binarisation method consists of training an end-to-end binary embedding extractor directly from aligned face images. Thus, the binarisation layer considers the loss function and is optimised for the task. We aim to obtain locality-preserving binary representations. The locality preserving property is defined by Equation (1) where a , p and n are three random points from the original space and $f(\cdot)$ is the projection function. The triplet loss function (shown in Equation (2)) is suitable for this task as the optimisation criterion is equivalent to Equation (1). To this end, we based our DNN on the FaceNet [7] architecture, which uses the triplet loss function for the training.

$$d(a, p) < d(a, n) \Rightarrow \|f(a) - f(p)\| < \|f(a) - f(n)\| \quad (1)$$

In the following subsections, we present the baseline face recognition system and describe the approaches taken to binarise the biometric data.

3.1 | Baseline face recognition system

Our goal is to obtain discriminating binary representations from faces that do not degrade the performance of the baseline system. The binarisation method proposed in this paper transforms Euclidean face embeddings into binary embeddings of different lengths. The Euclidean embeddings are

constructed using a deep neural network based on FaceNet [7]. In Ref. [24] we describe in detail the methodology we followed to create the face recognition system based on the OpenFace implementation [3]. We trained a convolutional DNN using the triplet loss function. The triplet loss function, given by Equation (2), takes a triplet comprised of an anchor x_i^a and a positive sample x_i^p from the same subject, and a negative sample x_i^n selected randomly from the rest of the data set. The training goal is to bring closer the anchor and positive samples and distance the negative sample using the margin α .

$$L = \sum_i^N \max\left(0, \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha\right) \quad (2)$$

The DNN architecture for training the face projection space is composed of 24 layers and 3 733 968 parameters. The training phase aims to obtain the best representation that separates the positive identities from negative ones using the triplet loss function. After the training phase, the network outputs a low dimensional representation of an input image consisting of a normalised Euclidean feature vector of size 128.

Figure 1 shows the pipeline of the face recognition system. First, the face is detected and aligned according to a predefined template. Afterwards, the aligned face is processed by the DNN in order to extract a Euclidean representation. This Euclidean representation constitutes the template that defines the user either for enrolment or verification.

Face alignment consists of three steps: face detection, landmark detection, affine transformation, and face cropping. The face detection is carried out using a deep convolutional neural network provided by OpenCV based on a Single-Shot-Multibox Detector (SSD) [25] and uses ResNet-10 architecture as a backbone. This model gives state-of-the-art performance with a low computational overhead. The image needs to be resized to 300×300 pixels to use the face detector. The image is provided in RGB format after subtracting the mean from each value. The output of the SSD detector is a bounding box. Given the face-bounding box, we use an implementation of Ref. [26] provided by DLIB [27] to detect the facial landmarks. Further details on the face alignment are provided in Ref. [28].

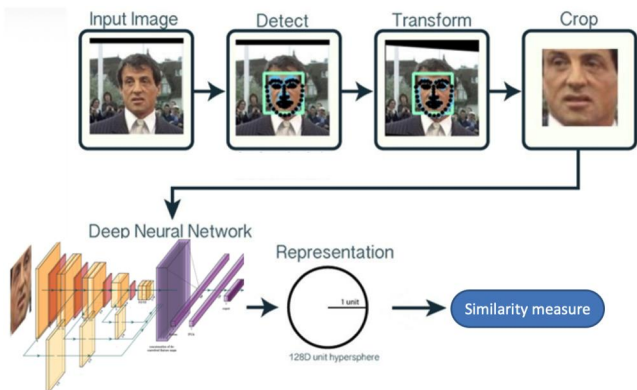


FIGURE 1 Pipeline of the baseline face recognition system

Finally, the Euclidean embedding extracted from the aligned face using DNN can be used for face recognition either in identification or verification scenarios. This paper aims to binarise the Euclidean embeddings with the least amount of degradation, which we explain in the next section.

3.2 | Locality preserving binary face representations using auto-encoders

Figure 2 shows the architecture of the proposed approaches. Both approaches (a) and (b) follow the same architecture. The difference lies in how the training data is used. In approaches (a) and (b), we opted to use an auto-encoder on top of the deep convolutional neural network (FaceNet based) to obtain the binary code.

The idea was to use an encoder to project the Euclidean representation that we get from the DNN onto another vector. This vector has the same size as the intended binary representation. Afterwards, we apply a custom binarisation layer on the vector and finally use a decoder to get back to the Euclidean representation.

The binarisation layer is defined as follows: In this layer, we apply a threshold to each input component. The output of this layer is defined in Equation (3). The input is compared to a threshold that is specified beforehand. The choice of the threshold is based on the type of the previous layer activation function. In our case, we chose a threshold of “0” as the previous activation function is the hyperbolic tangent. This layer does not have trainable parameters. In the back-propagation phase of the training, this layer is treated as the identity function, and its gradient is equal to 1.

$$F(input) = \begin{cases} 0, & \text{if } input \leq \text{threshold} \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

This idea has two benefits. First, we get more control over the length of the binary representation (we only need to modify the auto-encoder). The second benefit is that we get a continuous output from the auto-encoder, allowing us to use standard optimisation methods in conjunction with the triplet loss criteria. Figure 2 illustrates the example where we use a code length of P. First, the image is fed to the DNN, and we extract a Euclidean representation of size 128. Next, encode it on a P-component real vector, which is, in turn, binarised.

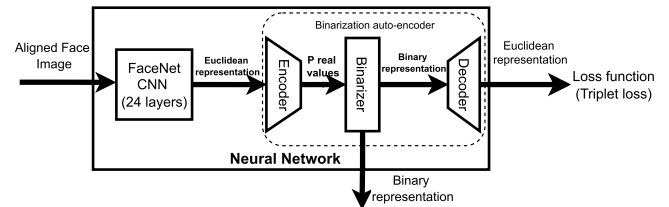


FIGURE 2 Block diagram of the binarisation method used in approaches (a) and (b). In approach (a), the whole model is trained from scratch. In approach (b), the FaceNet CNN is pretrained using the MS-celeb-1M

Then we reconstruct the initial Euclidean representation. The architecture described in Figure 2 is only used during the training phase. After training, we remove the decoder and obtain a binary code given a face image.

To put this idea into practice, we first needed to find an auto-encoder architecture suitable for the output. In other words, we sought to find the hyperparameters of the auto-encoder (number of hidden layers, width of the layers, and the activation functions) that result in the least degradation of the recognition performance compared to the original Euclidean representation. In this step, we did not use the binarisation layer. As the binarisation step generally degrades the performance, we would not be able to say whether the performance was degraded due to the auto-encoder or the binarisation step. The architecture that resulted in the least degradation was constructed using three layers. The encoder consisted of two linear layers with a hyperbolic tangent as an activation function. We used a single layer with a ReLU activation function for the decoder. The auto-encoder choice was based on the difference between the auto-encoder performance and the baseline performance of the original DNN architecture, which is 97.52% on the LFW. Once the auto-encoder architecture is set, we introduce the binarisation layer between the encoder and the decoder. The Final DNN architecture is presented in Table A1 of the appendix.

The difference between approach (a) and approach (b) is that in (a), we train the whole architecture from scratch, while in (b), we use a pretrained model on MS-celeb-1m. This model is described in Ref. [24]. Compared to approach (a), where the training is done from scratch, it is much faster for the DNN to converge towards good results. The loss of the models constructed using approach (a) stabilises around 1000 epochs compared to 100 epochs for models constructed using approach (b).

The following section presents the databases used for training and validating the models.

4 | DATABASES

We used the following databases to train and validate our models. We chose MS-celeb-1M for training the models because it is the biggest public data set for face recognition. In addition, the triplet loss function requires a high number of subjects with multiple images. LFW was chosen to evaluate the system's performance because it serves as the benchmark for most face recognition systems.

As for MOBIO, the data set was captured under challenging realistic use-case conditions, a person accessing his/her computer/phone. In this section, we provide a brief description of the particularities of these databases.

4.1 | Microsoft MS-celeb-1M

The MS-celeb-1M [1] is one of the largest publicly available databases. It has 100 k subjects and almost 10 M images. Popular search engines are used to provide about 100 images

for each subject. The images are collected based on their metadata, not their content. This results in the data set having a considerable amount of noise. The data set is constructed by Microsoft and is available for non-commercial use. Ref. [1] further describes the process of assembling the images and the metric used for the choice of the 100 K celebrity provided in the data set. We used the whole data set for training the neural network. The MS-celeb-1M database contains a significant portion of mislabelling because it was collected automatically using web crawlers.

In order to improve the performance, we leveraged clustering algorithms to clean the database. First, we applied Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [29] to reduce the mislabelling of the database. We worked under the assumption that there is no overlap between the identities of the labels provided in the database metadata. In other words, we can find multiple identities under the same label, but there is no overlap between the identities belonging to different labels. As the number of the identities in each label is unknown, we proceed by applying the DBSCAN clustering algorithm onto each label. The clustering is done on the embeddings computed using our model from Ref. [24]. The cluster with the highest number of samples is kept, and the remaining clusters are discarded. In cases where the number of samples in the most significant cluster is lower than three, the label is discarded.

Furthermore, the MS-celeb-1M database is biased towards the LFW data set as there is an overlap of the identities between both databases. We detail in Ref. [28] how we tried to reduce the bias towards the LFW database.

The cleaning reduced the training database to 80 k identities from the 100 k users provided in the MS-celeb-1M and reduced the total number of images from 10 to 4.5 M. This resulted in better overall performance for the baseline face recognition system. For example, in the case of the LFW database, using the same hyperparameters, the accuracy is improved from 97.53% to 98.82%. The impact of the cleaning is further shown in the case of the MOBIO database, where the accuracy of the baseline system improved from 90.6% to 98.9%.

4.2 | Labelled faces in the wild

The LFW data set contains 13233 target face images with considerable variability in facial expressions, age, race, occlusion, and illumination conditions. 1680 of the people pictured have two or more distinct photos in the data set. The only constraint on these faces is that they were detected by the Viola-Jones face detector [30]. The protocol specifies two views of the data set. View one is for model selection and algorithm development. It contains two sets: 1100 pairs per class (matched/mismatched) for training and 500 pairs per each class for testing. View 2 is designed for performance reporting. It is divided into 10 sets (folders), each with 300 matched pairs and 300 mismatched pairs. The cross-validation evaluation can be adopted among these 10 folders. The final

verification performance is reported as the mean recognition rate and standard error over the 10 fold cross-validation. It has to be noted that the task is to do pair matching: given a pair of images, the goal is to decide whether they belong to the same subject. This task is similar to face verification, except that the evaluation metrics proposed by the database collectors is the accuracy of the pair matching.

4.3 | MOBIO

The MOBIO database [31] is a bimodal (face/speaker) database recorded from 152 people. The database has a female-male ratio of nearly 1:2 (52 females 100 males). In total, 12 sessions were captured for each individual. It consists of three sets: training, development, and evaluation. In our experiments, we used only the development and evaluation sets.

The development set contains 42 subjects: 24 males and 18 females. As for the evaluation set, it comprised 58 subjects, 38 males and 20 females. Each subject has at least 120 videos. In the original protocol of the MOBIO database, the performance on the development set is measured using the equal error rate (EER) and the half total error rate (HTER) on the evaluation set. The evaluation protocol is described in Ref. [32]. The results are reported separately for males and females because separating males from females for speaker recognition gives better results. Therefore face recognition experiments follow the same principle in this protocol.

However, in our case, we applied the 10-fold cross-validation pair-matching protocol similar to the LFW database to have the same evaluation metric for both databases. We concatenated the development and evaluation partitions to obtain a single testing partition composed of 100 subjects (62 males and 28 females). We use three frames from each video. Frames where the face is not present, are discarded. In order to have a balanced accuracy, we used 50 000 matched pairs and 50 000 mismatched pairs. The accuracy is computed on the 100 000 pairs using 10-fold cross-validation. The performance using the original protocol is reported in the appendix in Figure A1 and Table A2.

5 | BIOMETRIC PERFORMANCE OF THE BINARY REPRESENTATIONS

In this section, we present the biometric performance of the binary representations. We evaluate the performance on the LFW and the MOBIO databases using the accuracy, as a common evaluation metric, computed using the 10-fold cross-validation protocol.

We evaluate the recognition performance and the entropy of the models. As the goal of the work is to binarise the biometric samples to be suitable for biometric crypto-systems and biometric protection schemes, the binary representations should have high entropy and good recognition performance.

In approach (a), we train the network, shown in Figure 2, from scratch on the MS-celeb-1M data set using the triplet loss

function. We report in Table 1 the performance of this approach for various lengths of the binary representations. The training was carried out for 1000 epochs. We note that the best performance on LFW is obtained with 512-bit representations. On the other hand, 512-bit representations provide the best performance on the MOBIO data set. We attribute that to the overlap of the original MS-celeb-1M data set with the LFW data set. As the representation length grows, the model overfits to MS-celeb-1M, resulting in worse performance on MOBIO.

When the length of the embeddings reaches 4096 bits, the recognition performance decreases dramatically. On LFW, the accuracy plummets from 93% to 81% compared to the representation with a length of 2048. The performance degradation is more accentuated on the MOBIO data set, where the error reaches almost 50%. We attribute the cause of the degradation when using 4096-bit embeddings to the loss of information in the training phase of the neural network due to the thresholding process. The information propagated backward is not enough to optimise the system's parameters. The number of trainable parameters in the auto-encoder evolves exponentially from 33024 parameters for representations with a length of 128 bits to 1 056 768 parameters for the 4096-bit representations.

Studying the biometric performance of the binary representation alone is not enough, especially when we are trying to have long representations. Appending a fixed portion to all the representations will not degrade the recognition performance of the system. However, as our primary goal is to obtain a long binary representation, we need to study the entropy of the representations. We report in Table 2 the entropy of the binary representations according to their length.

The entropy is measured on 5 million samples from MS-celeb-1M. We use Monte Carlo random sampling in order to compute the entropy. From the 5 M samples, we select 500 k samples randomly and measure the entropy based on those 500 k samples. This step is repeated for 1000 iterations. The

TABLE 1 Impact of the length of the binary representations on the biometric performance of approach (a): Training the auto-encoder using triplet loss from scratch

Length	Accuracy on LFW %	Accuracy on MOBIO %
Baseline system	97.52	90.58
128*(median)	89.32	79.74
128	91.73	82.50
256	93.18	83.50
512	94.12	84.23
1024	93.62	81.46
2048	93.07	79.46
4096	81.13	53.60

Note: The baseline system is the system used in [17]. The results in the second row (row '128*') are obtained by applying a median binarisation on the output of the CNN used in Ref. [17]. The maximum standard deviation (std) on Labelled Faces in the Wild (LFW) is around 1%. The maximum std on MOBIO is around 0.1%. Best results are presented in BOLD.

TABLE 2 Entropy of the representations created using approach (a)

Length	$p(x=1)$	Entropy
128	0.487	98.26
256	0.532	113.87
512	0.514	163.4
1024	0.496	116.65
2048	0.511	143.92
4096	0.503	49.87

Note: The entropy was measured using 5 M samples from MS-celeb-1M. $p(x=1)$ is the probability of a bit is equal to 1. Best results are presented in BOLD.

entropy provided in the tables is the average of the 1000 iterations. Representations of length 512 provide the highest entropy with 163 bits. On the other hand, embeddings of length 4096 give the lowest value for entropy, which is consistent with their biometric recognition performance.

The results of the approach (a), especially the low entropy, led us to use the pretrained face recognition models instead of training from scratch. Table 3 reports the performance of the system when we use a pretrained CNN. Using a pretrained CNN significantly improves performance, especially for embeddings with 4096 bits. In addition, the pretraining reduces the loss of information introduced by the auto-encoder. Using a pretrained model on the cleaned version of MS-celeb-1M and adding the auto-encoder previously discussed resulted in better biometric verification performance compared to training the model from scratch. The pretrained CNN is the FaceNet model trained on the same data set as the auto-encoder. So, when we present the performance of the models trained on the original/cleaned version of MS-celeb-1M, the pretrained CNN is trained separately on the same set as the whole module.

We report the entropy of the binary representations obtained using an auto-encoder with a pretrained CNN in Table 4. For the version trained on the original MS-celeb-1M, we see that the entropy of the keys reaches its maximum of 260 for representations of size 1024. Besides, the $p(x=1)$ is around 0.5 (except for length 4096), which shows that many bits of the representations are constant. In addition, when we use the cleaned training database for training the system, we see that entropy improves significantly, in particular when the length of the representation exceeds 512 bits.

To estimate the degradation of the biometric performance introduced by the binarisation, we compare the performance of the approach (a) to the system presented in Ref. [24]. For approach (b), we compare the performance of the binarised embeddings to the pretrained CNN that was used. Approach (a) shows higher degradation of the performance, from 97.53% to 94.12% accuracy on LFW and from 90.58% accuracy on MOBIO to 84.23%. The degradation is more pronounced on the MOBIO database due to the bias in the original version of MS-celeb-1M towards the LFW data set.

As for approach (b), we present two cases. The first case is when the pretrained CNN and the auto-encoder are trained on the original MS-celeb-1M. In this case, as shown in Table 3, the

TABLE 3 Impact of the length of the binary representation on the biometric recognition performance of approach (b) (using a pretrained CNN with an auto-encoder)

Length	Accuracy on LFW %		Accuracy on MOBIO %	
Pretrained CNN	97.52	99.22	90.58	98.93
128*(median)	89.32	93.22	79.74	90.15
128	94.88	97.30	81.31	95.27
256	95.37	97.50	87.62	97.84
512	95.85	98.80	87.11	98.28
1024	96.32	99.12	89.35	98.87
2048	95.06	99.00	85.60	98.58
4096	95.15	99.00	80.12	98.90

Note: Values in bold are given by models trained using the cleaned version MS-celeb-1M. The first row is provided to show the degradation of recognition performance between the initial system (Euclidean embeddings) and the binarised embeddings. By 'pretrained CNN', we denote the initial OpenFace DNN. The results in the second row (row '128*') are obtained by applying a median binarisation on the output of the pretrained CNN.

TABLE 4 Entropy of the representations created using approach (b)

Length	$p(x=1)$	Entropy		
128	0.497	0.489	112.22	116.20
256	0.486	0.481	205.67	233.59
512	0.493	0.482	252.01	473.74
1024	0.506	0.454	261.29	944.24
2048	0.498	0.315	223.99	1679.25
4096	0.826	0.308	179.08	3349.47

Note: The entropy was measured using 5 M samples from MS-celeb-1M. $p(x=1)$ is the probability of a bit being equal to 1. Values in bold are given by models trained using the cleaned version MS-celeb-1M.

accuracy on LFW is decreased by about 1%–2% compared to the baseline. On the other hand, the accuracy on the MOBIO data set improved compared to the performance of approach (a). We attribute the difference of behaviour of the system to the overlap between the training and LFW databases. However, when the training is carried out on the cleaned database, the degradation on both data sets is lower than 1%. On the LFW database, we obtain 99.12% accuracy using the binary representations, whereas we get 99.22% accuracy using the baseline system. The same applies to the MOBIO database, where we get 98.9% accuracy using the binary representations compared to an accuracy of 98.93% with the baseline system. This shows that our binarisation methods are highly dependent on the quality of the training data. By the quality of the training data, we refer to the level of the noise, mislabelling, quality of the images, and size of the database. If we have little data, it will result in low entropy of the representations. The mislabelling and noise will also reduce the system's accuracy and lower the entropy of the representations at the same time. As shown in Table 4, the entropy of the representations depends on the

TABLE 5 Performance of the classical binarisation methods on the Labelled Faces in the Wild (LFW) data set

Encoding	Length (bits)	Accuracy on LFW (%)	Entropy
Euclidean representation (OpenFace)	128 floats	99.22	~
DBR	256	97.28	253.23
	1024	84.25	650.50
BRGC	256	97.37	146.04
	1024	96.17	561.74
LSSC	348	97.38	148.60
	1024	98.62	409.03
Sparse	512	96.93	275.31
	1024	94.35	418.67
Ours	1024	99.12	944.24

Note: The binarisation methods are applied to the output of our version of OpenFace CNN trained on the **cleaned** version of MS-celeb-1M. The entropy of the methods is computed using the same approach presented previously. Best results are presented in BOLD.

quality of the training data set (non-cleaned/cleaned). For the non-cleaned version, representations with lengths longer than 256 bits have no further useful information. As for the cleaned version, this behaviour appears when we exceed the length of 4096 bits. The proposed auto-encoder can provide longer representations, but their real length, which is shown through their entropy (See Table 2)

Finally, in both approaches (a) and (b), the performance is better than binarising simply using the median as described in Ref. [33]. Moreover, our method has the advantage of providing arbitrary length representations limited only by the quality of the training data set. The representation length can thus be adapted to the sensitivity of the application.

We present in Table 5 a comparison between our proposed approach and some classical binarisation methods. These classical methods were presented in Refs. [9, 16] and benchmarked on the AR and FERET data sets. We followed the proposed approach presented in Ref. [9] for binarising the output of the CNN by quantising the feature space and applying an encoding to the codebook obtained in the quantisation step. We follow the same processing chain presented in Ref. [9], but we used our DNN features as input for the binarisation methods. The binarisation methods that we re-implemented are the following:

- Direct Binary Representation (DBR), where the decimal values from the quantisation are directly convected into their binary representations.
- Binary Reflected Grey Code (BRGC), similar to the DBR method, where the decimal values are encoded directly to binary format using their BRG representations.
- Linearly separable Subcode (LSSC) [16], an encoding method that aims to keep the distance from the decimal space to the binary space.
- Sparse, in this scheme, which is similar to one-hot encoding, the number of encoded bits per real value is equal to the number of quantisation intervals, and only one bit is set to one per encoding.

We followed an equal-width quantisation approach where the feature space is divided into intervals of the same size.

In our comparison, we used the same output lengths for each of the systems as in Ref. [9]. Furthermore, we also adapted the schemes to obtain 1024 bits for all the methods, mainly by changing the number of quantisation intervals. For example, for DBR to obtain representations with 1024 bits, we used 256 quantisation intervals to obtain a DBR representation on 8 bits for each real value. BRGC, LSSC, and Sparse were quantised over 256, 9, and 8 intervals, respectively.

As shown in Table 5, our approach gives better recognition performance than the classical methods. Furthermore, the entropy of our approach is higher than the classical approaches presented. For example, LSSC shows the best performance among the studied classical binarisation approaches with 98.62% accuracy on LFW compared to the original baseline of 99.2%. Thus, the recognition degradation of this approach is minor. However, it provides less than half the entropy provided by our binarisation approach. In addition, some of the methods show significant degradation of the performance when using longer representations (such as DBR) and, as such, limiting the length of the representation. BRGC and Sparse, and especially DBR, suffer from performance degradation when increasing the length of the representations. We attribute the degradation of the performance for DBR to two factors: first, the high number of quantisation intervals; second, the fact that the DBR code does not conserve distances as opposed to LSSC and BRGC. On the other hand, our method keeps the system's performance even with much longer representations as we do not need to change the number of quantisation intervals by increasing the number of neurons in the bottleneck layer in the auto-encoder; we can increase the length of the binary representation.

In the following section, we provide a use case of the binary representations consisting of a cancellable face verification system.

6 | APPLICATION TO CANCELLABLE BIOMETRICS

Biometrics systems are strongly associated with identity, and therefore, biometric recognition creates a strong link between the user's identity and the authenticator. However, many privacy concerns are being raised about biometrics. Since biometric characteristics are permanently associated with the person, they cannot be replaced in case of compromise. This lack of revocability is a serious issue for user authentication systems. Moreover, biometric templates originating from the same biometric characteristics stored in different databases are similar. Therefore, biometrics lack diversity, and two biometric databases can be cross-linked, compromising the user's privacy. Recovery of biometric data from the biometric references and possibly revealing physical conditions bring additional privacy issues with biometric systems.

Cancellable biometrics is proposed in order to address these problems. It consists of transforming the original biometric template to obtain a cancellable biometric reference that can be revoked. Therefore, when a biometric template is compromised, it can be cancelled and replaced.

6.1 | Cancellable system requirements

There are some main criteria that a cancellable biometric template should satisfy:

- **Performance:** the cancellable biometric system should not degrade the verification performance of the underlying baseline biometric system.
- **Revocability:** if the protected biometric template is stolen, it should be possible to revoke that template and reissue a new one;
- **Diversity:** is the maximum number of independently protected templates that can be created from one biometric sample.
- **Irreversibility:** it should be computationally infeasible to obtain the original biometric template from the protected template.
- **Unlinkability:** the protected biometric templates created from the same biometric sample using two different secret keys should not be linkable.

In the following subsection, we present and evaluate the performance of the biometric protection scheme applied to the binary representations created using our binarisation method. In the following evaluation, we use the terminology of the ISO/IEC 24745:2011 [34]. We use **PI** to denote the Pseudonymous Identifier and **SD** for Supplementary Data.

6.2 | Proposed cancellable system

To protect the template, we apply the shuffling scheme proposed by Kanade et al. in Ref. [35]. The shuffling scheme

(shown in Figure 3) uses a binary shuffling key. Since this key is a long bit-string, it is stored on a secure token, or it can be derived from a password. The binary embedding is divided into blocks of the same length. Two distinct parts are created: the first part contains all the blocks corresponding to the positions where the shuffling key bit value is '1'. All the remaining blocks are taken into the second part. These two parts are concatenated to form the shuffled binary embedding, treated as the protected template. The original and shuffled templates have a one-to-one correspondence. A block from the original vector is placed at a different position in the shuffled embedding. When two binary embeddings are shuffled using the same shuffling key, the absolute positions of the blocks change, but this change occurs in the same way for both of the representations. As a result, the Hamming distance between them does not change. On the other hand, if they are shuffled using two different keys, the result is a randomisation of the representations, and the Hamming distance increases.

For this use case, we chose a block size of '1' compared to '7' in Ref. [35]. This has two main advantages. First, the size of the shuffling key will be longer, thus harder to brute-force. Secondly, the permutation space becomes bigger, allowing for a higher number of possible templates. The shuffled binary embedding, which is the cancellable template, is the result of combining the biometric sample and the Supplementary Data (**SD**) (the shuffling key in our case). Therefore, it can be revoked in case of compromise, and a new template can be generated by changing the shuffling key. In our case, we chose a block size of "1" with a shuffling key of size 1024. The shuffling keys can be either generated and stored in the Secure Element or derived from the password using, for example, a password-based key derivation function such as PBKDF2.

According to the results reported in Table 6, a binary embedding of 1024 bits gives the best trade-off between size and performance. As such, all subsequent evaluation analyses are carried out using 1024-bit representations.

6.2.1 | Biometric recognition performance

The performance of the verification system is an important point that must not be degraded by the transformation scheme. Therefore, for a fair comparison, first, the biometric verification performance of the baseline biometric system should be evaluated, then the performance of the proposed cancellable biometric system. It is necessary to evaluate the system

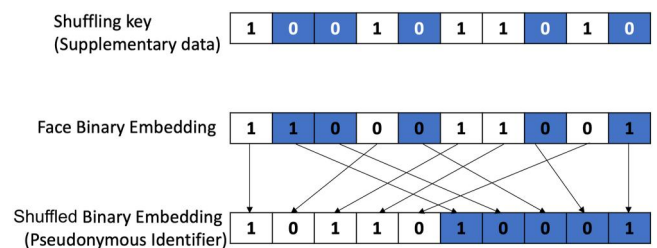


FIGURE 3 Shuffling scheme with block size of '1' bit

TABLE 6 Impact of the length of the shuffled binary representations obtained following approach (b) (using a pretrained CNN with an auto-encoder) on the recognition performance

Length	Accuracy on LFW %		Accuracy on MOBIO %	
128*	98.32	98.00	99.72	99.67
128	98.27	98.82	99.88	99.67
256	99.68	99.22	99.91	99.88
512	100	99.80	100	100
1024	100	99.88	100	100
2048	100	99.88	100	100
4096	100	99.77	100	100

Note: Values in **bold** are given by DNN models trained using the cleaned version MS-celeb-1M. The results in the second row (row '128*') are obtained by applying a median binarisation on the output of the initial OpenFace DNN.

performance when one of the two factors is compromised. Hence, two impostor scenarios are considered:

- Stolen biometric data: when the biometric data for the user is compromised. Here, an impostor will try to provide the stolen biometric data with the wrong **SD**;
- Stolen Supplementary data: when the **SD** of the user is compromised. Here, an impostor will try to provide erroneous biometric data with the stolen **SD**.

The biometric recognition performance of the system is reported in Table 6. The performance of the system is improved compared to using non-shuffled representations. Moreover, we obtain better overall performance for the shuffling when using our proposed binarisation method compared to using median threshold as shown in the first and second row of Table 6. We also note that thanks to the fact that we can control the length of the generated binary representation, we can improve the recognition performance by using longer representations.

For the stolen biometric scenario, the system has a False Acceptance Rate (FAR) of 0%. This point is further developed in the unlinkability analysis. Therefore, the protected biometric templates created from the same biometric sample using two different secret keys should not be linkable, which is the same as using a compromised biometric sample with a different key.

As for the stolen **SD** scenario, the performance of the system reverts to the case of non-shuffled representations shown in Table 3.

6.2.2 | Diversity

It is necessary to calculate the maximum number of pseudonymous identifiers (a pseudonymous identifier (**PI**) is a part of a renewable biometric reference that represents an individual or data subject) that can be generated. After that, unthinkability and irreversibility analysis should be done as a function of **PI** issued. In the case of the previously described

shuffling scheme, the maximum number of **PI** is given using the number of possible permutations. Moreover, because the decision-making is based on a threshold comparison, we should not account for templates falling in the same neighbourhood. We estimate the maximum number of templates using the Hamming-packing bound. Using a threshold $t = 0.2$, for binary representations of length 1024, we get around 2^{194} possible **PI** for each user.

$$\begin{aligned} \text{Number Of PI} &= \frac{\text{number of permutation}}{\text{volume of Hamming spheres}} \\ &= \frac{1024!}{512! \sum_{k=0}^{t \times 1024} \binom{1024}{k}} \approx 2^{194} \quad (4) \end{aligned}$$

6.2.3 | Irreversibility

There are two types of irreversibility analysis. The first type is to analyse whether we can revert to the original template given the **SD**. The second analysis is the analysis of the protected template without having the **SD**. As the applied transformation is a shuffling of the bits of the embedding without a loss of information, given the second factor, the scheme is fully reversible. However, without access to the second factor and prior knowledge about the distribution of the non-shuffled templates, it is computationally not feasible to revert to the original binary embedding as the number of permutations to be tested, which is equal to $\frac{1024!}{512! 512!} \approx 2^{1018}$ is too big to be brute-forced.

6.2.4 | Unlinkability

For this metric, we follow the methodology defined in Ref. [36]. Two types of score distributions will be analysed for the assessment of the unlinkability provided by the protected templates:

- **Mated instances:** scores computed from templates extracted from different samples of the same subject using different keys.
- **Non-mated instances:** scores obtained from templates generated from samples of different subjects using different keys.

As described in Ref. [36], two measures are computed, $D_{\leftrightarrow}(s) \in [0,1]$ gives an estimation of the linkability of a system for a specific score s , and $D_{\leftrightarrow}^{ys} \in [0,1]$ gives an estimation of the linkability of a system as a whole, independently of the score. If for a specific score s_0 $D_{\leftrightarrow}(s_0) = 0$, this means that the system is fully unlinkable for this particular score. Also, if $D_{\leftrightarrow}^{ys} = 0$ where both score distributions (mated and non-mated) are overlapping, this means that the system is fully unlinkable for the whole score range. The computation of $D_{\leftrightarrow}(s)$ and D_{\leftrightarrow}^{ys} depends on the prior probability ratio ω of the mated and non-mated distributions, which may result in $D_{\leftrightarrow}^{ys} = 0$ even if the

distributions are not perfectly overlapping. In our case, the prior probability ratio is $\omega = 0.2$. Using this value for ω the distribution of mated and non-mated scores overlap, thus making the function $D_{\leftrightarrow}(s)$ identically zero over the range of the possible scores. In addition D_{\leftrightarrow}^{ys} is equal to 0, rendering the system fully unlinkable. The scores used to estimate the probabilities are computed using the whole LFW data set of 5749 users. For each user, we generate 50 different shuffling keys and thus 50 protected templates. By considering the whole population of the LFW data set, we get around 14 M mated scores and 80 000 M non-mated scores. To have the same number of samples from each population, we sample uniformly 10 M mated scores and non-mated scores. Hence, $D_{\leftrightarrow}(s)$ and D_{\leftrightarrow}^{ys} are estimated in the mean case and do not take account of user-specific distributions. Based on $D_{\leftrightarrow}(s)$ and D_{\leftrightarrow}^{ys} we conclude that the proposed system is fully unlinkable for the whole score range.

To further study the generalisation of the unlinkability of the system, we study the unlinkability metric for $\omega = 1$. Figure 4 shows this case where we obtain a global linkability measure of $D_{\leftrightarrow}^{ys} = 0.03$. This is due to the fact that, for similarity scores $s < 0.48$, it is more likely that templates stem from mated instances. However, since the probability of obtaining such scores is very low, the system is almost fully unlinkable, hence the low value for D_{\leftrightarrow}^{ys} .

The diversity, irreversibility, and unlinkability metrics are tightly correlated. If the system cannot satisfy the diversity requirement, and as such, cannot create different **PI**s using the same biometric data with different **SD**s, then the identities will be linkable. Furthermore, if the irreversibility requirement is not satisfied, the templates can be linked. Finally, if the system is fully linkable, then it does not satisfy the diversity requirement as all the generated **PI**s are equal. Furthermore, even if the system is not fully linkable

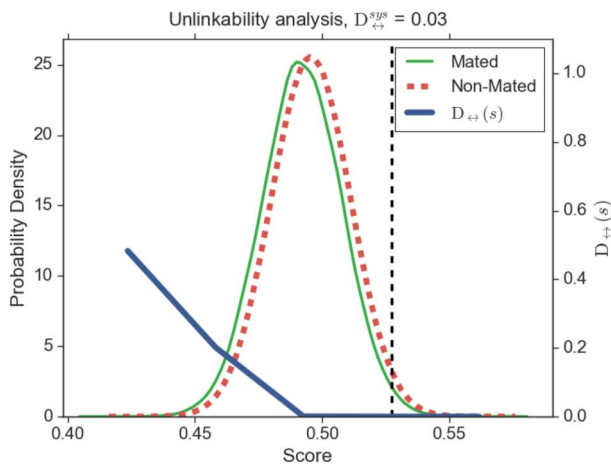


FIGURE 4 Unlinkability analysis of the system based on scores computed on the Labelled Faces in the Wild (LFW) data set for $\omega = 1$. Templates used are of length 1024. The templates are obtained using DNN, created corresponding to approach (b) (using a pretrained CNN with an auto-encoder) and trained on the **cleaned** version of MS-celeb-1M

and only partially unlinkable, it will result in easier attacks on the original templates.

In addition to the evaluation criteria proposed by the ISO/IEC 24745:2011 standard [34], in the case of cancellable biometrics, one should check if the security of the system is only based on the second factor. Cancellable systems tend to rely on the second factor ignoring the biometric component, which is one of the shortcomings of cancellable biometrics as shown in Refs. [37, 38]. In fact, for the used shuffling scheme, if all the users have the same initial representation, after shuffling, we obtain 100% verification accuracy. Thus, the protection scheme based on shuffling benefits greatly from the security of the second factor. However, the combination of the binarisation method we propose with the shuffling scheme constitutes a system that relies on biometrics as well as on the second factor. This is especially shown in the difference between the systems trained on the original and cleaned version of MS-celeb-1M. The degradation of performance of the cancellable system shown in Table 6 when the training is done on the cleaned version of the MS-celeb-1M is, in fact, due to the bad quality of the images used in the tests. The system should not accept these images because the face is obstructed, distorted, or not present. When the binary embedding extractor is trained on the cleaned data set, the system rejects client–client tests where either the enrolment or probe samples are of low quality. On the other hand, the version trained on the original version of MS-celeb-1M (non-cleaned) accepts these images because the verification is done using the second factor, not the biometric reference. Examples of the images with bad quality are presented in Figure 6b. The test scores from these images are circled in red in Figure 5 (Hamming Distance > 0.4). The face image samples are taken from the MOBIO data set. The images of bad quality, such as those presented in the figure, are not accepted by the cancellable system based on binarised DNN embeddings

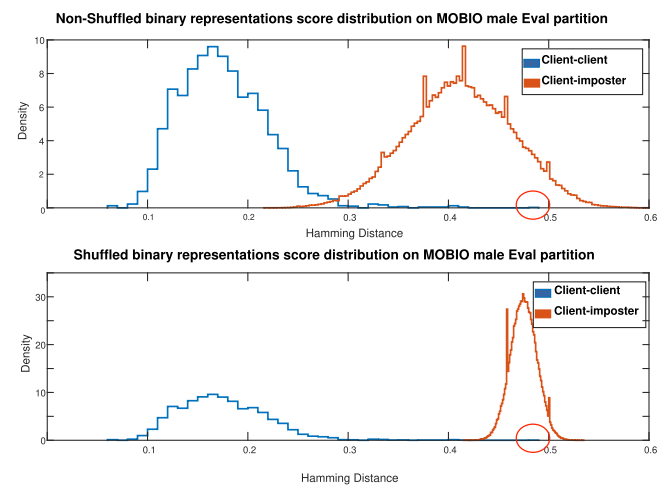


FIGURE 5 Impact of the shuffling on the score distribution of the data. Score distribution from templates of length 1024. The templates are obtained using the DNN corresponding to approach (b) (using a pretrained CNN with an auto-encoder) and trained on the **cleaned** version of MS-celeb-1M

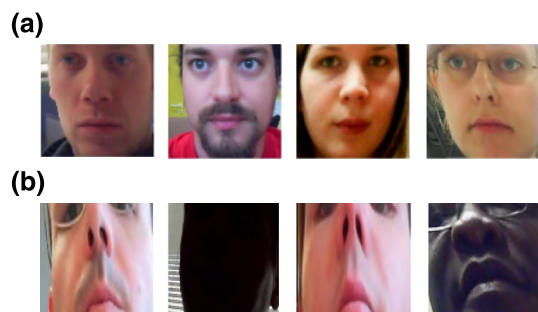


FIGURE 6 Face image samples taken from the MOBIO database. Face detection is done using the OpenCV Single-Shot-Multibox Detector (SSD) face detector. Alignment is done using the DLIB 68 points landmark detector. (a) Examples of face images with good quality (b) Examples of face images with bad quality

trained on the **cleaned** version of MS-celeb-1M. The system is intended to work with images such as those in Figure 6a.

This shows that the system considers the biometric information and does not only focus on the second factor. As the system trained on the cleaned version of MS-celeb-1M rejects images of the same user of low quality, it does not rely solely on the second factor.

7 | CONCLUSIONS

This paper presents a novel approach to extract binary embeddings directly from face images using a deep neural network. We followed a data-driven approach to binarise the embeddings based on using auto-encoders under supervised training with the ‘Triplet loss’ loss function.

The binary embeddings are analysed in terms of biometric recognition performance and entropy. The performance is evaluated on the LFW and MOBIO databases. The degradation of performance on both databases is around 0.1%. We obtain 99.12% accuracy on the LFW database, using the binary representation, compared to 99.22% accuracy using the baseline system. The same applies to the MOBIO database, where we get 98.90% accuracy using the binary embeddings compared to an accuracy of 98.93% of the baseline system. Using DNN to extract the binary embeddings results in representations with high entropy and high recognition performance. Compared to the baseline Euclidean representations, the proposed binary embeddings give a state-of-the-art performance on both databases with almost negligible degradation.

The approach proposed in this paper can be applied to any continuous representation, not only Euclidean face representations. Moreover, the binarisation technique constitutes a locality-preserving hash where the relative distance between the input values is preserved in the relative distance between the output hash values. The representation can be used for multiple applications such as similarity search, database search, and biometric systems.

Furthermore, the binarisation method provides arbitrary length representations that are limited only by the quality of the

training database. The embedding length can thus be adapted to the sensitivity of the application. In addition, we compared our binarisation approach to some classical binarisation methods presented in Ref. [9] and show that our method has better biometric recognition performance and higher entropy than the presented methods.

The binary embeddings are also used to create a cancellable face recognition system based on a shuffling transformation using a second factor. The cancellable system is analysed according to the standardised metrics given by the ISO/IEC 24745:2011. We show that the cancellable system gives high accuracy and unlinkable templates when the second factor is not compromised. When the second factor is compromised, the system's security is assured by the recognition performance of the binary representations, which is comparable to the baseline non-binarised system. Furthermore, the quality of the binary representations impacts the behaviour of the cancellable system. If the discriminative power of the representations is low, the cancellable system depends mainly on the second factor, which results in higher FAR.

These representations are meant to be used in a crypto-biometric key regeneration scheme based on fuzzy commitment. This is why we seek to obtain long binary representations with high entropy.

ACKNOWLEDGEMENTS

This work was partially supported by the SpeechXRays and Empathic projects that have received funding from the European Commission's Horizon 2020 research and innovation program, under Grant Agreements Nos. 653586 and 769872.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study.

ORCID

Mohamed Amine Hmani  <https://orcid.org/0000-0002-2403-2643>

REFERENCES

- Guo, Y., et al.: Ms-celeb-1m: a dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision, pp. 87–102. Springer (2016)
- Huang, G.B., Learned-Miller, E.: Labeled faces in the wild: updates and new reporting procedures. Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep, pp. 3–14 (2014)
- AMOS, B., Ludwiczuk, B., Satyanarayanan, M.: Openface: A General-Purpose Face Recognition Library with Mobile Applications. CMU Sch. Comput. Sci. 6(2). 20 (2016)
- Deng, J., et al.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4690–4699 (2019)
- Erik, L.-M., et al.: LFW: Results (2019)
- Liu, W., et al.: SphereFace: deep hypersphere embedding for face recognition. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua, pp. 6738–6746 (2017)

7. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823 (2015)
8. Lim, M., Teoh, A.B.J., Kim, J.: Biometric feature-type transformation: making templates compatible for secret protection. *IEEE Signal Process. Mag.* 32(5), 77–87 (2015). <https://doi.org/10.1109/msp.2015.2423693>
9. Drozdzowski, P., et al.: Benchmarking binarisation schemes for deep face templates. In: *Proceedings - International Conference on Image Processing*, pp. 191–195. *ICIP* (2018)
10. Bringer, J., Despiegel, V.: Binary feature vector fingerprint representation from minutiae vicinities. In: *IEEE 4th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2010*, pp. 1–6 (2010)
11. Cappelli, R., Ferrara, M., Maltoni, D.: Minutia Cylinder-Code: a new representation and matching technique for fingerprint recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 32(12), 2128–2141 (2010). <https://doi.org/10.1109/tpami.2010.52>
12. Chen, C., Veldhuis, R.: Binary biometric representation through pairwise adaptive phase quantization. *EURASIP J. Inf. Secur.* 2011(1), 543106–16 (2011). <https://doi.org/10.1155/2011/543106>
13. Chen, C., et al.: Biometric quantization through detection rate optimized bit allocation. In: *Eurasip Journal on Advances in Signal Processing*, 2009, pp. 1–16 (2009)
14. Kevenaar, T.A., et al.: Face recognition with renewable and privacy preserving binary templates. In: *Proceedings - Fourth IEEE Workshop on Automatic Identification Advanced Technologies, AUTO ID 2005*, pp. 21–26 (2005)
15. Lee, H., et al.: A secure biometric discretization scheme for face template protection. *Future Generat. Comput. Syst.* 28(1), 218–231 (2012). <https://doi.org/10.1016/j.future.2010.11.006>
16. Lim, M.H., Teoh, A.B.J.: A novel encoding scheme for effective biometric discretization: linearly separable subcode. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(2), 300–313 (2013). <https://doi.org/10.1109/tpami.2012.122>
17. Schlett, T., Rathgeb, C., Busch, C.: A binarization scheme for recognition based on multi-scale block local binary patterns. In: *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–4. *IEEE* (2016)
18. Chen, L., et al.: Face template protection using deep LDPC codes learning. *IET Biom.* 8(3), 190–197 (2018). <https://doi.org/10.1049/iet-bmt.2018.5156>
19. Mai, G., et al.: SecureFace: face template protection. *IEEE Trans. Inf. Forensics Secur.* 16, 262–277 (2021). <https://doi.org/10.1109/tifs.2020.3009590>
20. Schlemper, J., et al.: Deep Hashing Using Entropy Regularised Product Quantisation Network, pp. 1–11 (2019)
21. Pandey, R.K., et al.: Deep secure encoding for face template protection. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 77–83 (2016)
22. Jindal, A.K., Chalamala, S., Jami, S.K.: Face template protection using deep convolutional neural network. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2018-June, pp. 575–583 (2018)
23. Carreira-Perpinán, M.A., Raziperchikolaei, R.: Hashing with binary autoencoders. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 557–566 (2015)
24. Hmani, M.A., Petrovska-Delacrétaz, D.: State-of-the-art face recognition performance using publicly available software and datasets. In: *2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pp. 1–6. *IEEE* (2018)
25. Liu, W., et al.: Ssd: single shot multibox detector. In: *European Conference on Computer Vision*, pp. 21–37. *Springer* (2016)
26. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874 (2014)
27. King, D.E.: Dlib-ml: a machine learning toolkit. *J. Mach. Learn. Res.* 10, 1755–1758 (2009)
28. Hmani, M.A., Mubaa, A., Petrovska-Delacrétaz, D.: Voice biometrics: technology, trust and securitychap. In: *Joining Forces of Voice and Facial Biometrics: A Case Study in the Scope of NIST SRE19*, pp. 187–217. *Security Institution of Engineering and Technology*. (2021)
29. Ester, M., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, vol. 96, pp. 226–231 (1996)
30. Viola, P., Jones, M., et al.: Rapid object detection using a boosted cascade of simple features. *CVPR* 1(1), 511–518, 3 (2001)
31. Mccool, C., et al.: Bi-modal person recognition on a mobile phone: using mobile phone data. In: *2012 IEEE International Conference on Multimedia and Expo Workshops*, pp. 635–640. *IEEE* (2012)
32. Günther, M., Shafey, L.E., Marcel, S.: Face recognition in challenging environments: an experimental and reproducible research surveypp. In *Face recognition across the imaging spectrum*, pp. 247–280. *Springer International Publishing*. Cham (2016)
33. Hmani, M.A., et al.: Evaluation of the H2020 SpeechXRays project cancelable face system under the framework of ISO/IEC 24745: 2011. In *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pp. 1–6. *IEEE* (2020)
34. Techniques, I.J.S.S.: Information Technology - Security Techniques - Biometric Information ProtectionInternational Organization for Standardization (2011). *ISO/IEC 24745:2011*
35. Kanade, S.G., Petrovska-Delacrétaz, D., Dorizzi, B.: Enhancing information security and privacy by combining biometrics with cryptography. *Synthesis Lectures on Information Security, Privacy, and Trust* 3(1), 1–140 (2012). <https://doi.org/10.2200/s00417ed1v01y201205spt003>
36. Gomez-Barrero, M., et al.: General framework to evaluate unlinkability in biometric template protection systems. *IEEE Trans. Inf. Forensics Secur.* 13(6), 1406–1420 (2018). <https://doi.org/10.1109/tifs.2017.2788000>
37. Kong, A., et al.: An analysis of BioHashing and its variants. *Pattern Recogn.* 39(7), 1359–1368 (2006). <https://doi.org/10.1016/j.patcog.2005.10.025>
38. Rathgeb, C., Uhl, A.: A survey on biometric cryptosystems and cancelable biometrics. *EURASIP J. Inf. Secur.* 2011(1), 3 (2011). <https://doi.org/10.1186/1687-417x-2011-3>

How to cite this article: Hmani, M.A., Petrovska-Delacrétaz, D., Dorizzi, B.: Locality preserving binary face representations using auto-encoders. *IET Biome.* 11(5), 445–458 (2022). <https://doi.org/10.1049/bme2.12096>

APPENDIX

TABLE A1 Details of the nn4.small2 Inception architecture, which is a version of the nn4 model from FaceNet [7] hand-tuned by Ref. [3] to have less parameters

type	output size	#1 × 1	#3 × 3 reduce	#3 × 3	#5 × 5 reduce	#5 × 5	Pool proj
Conv1 ($7 \times 7 \times 3, 2$)	$48 \times 48 \times 64$						
Max pool + norm	$24 \times 24 \times 64$						m $3 \times 3, 2$
Inception (2)	$24 \times 24 \times 192$		64	192			
Norm + max pool	$12 \times 12 \times 192$						m $3 \times 3, 2$
Inception (3a)	$12 \times 12 \times 256$	64	96	128	16	32	m, 32p
Inception (3b)	$12 \times 12 \times 320$	64	96	128	32	64	l_2 , 64p
Inception (3c)	$6 \times 6 \times 640$		128	256,2	32	64,2	m $3 \times 3, 2$
Inception (4a)	$6 \times 6 \times 640$	256	96	192	32	64	l_2 , 128p
Inception (4e)	$3 \times 3 \times 1024$		160	256,2	64	128,2	m $3 \times 3, 2$
Inception (5a)	$3 \times 3 \times 736$	256	96	384			l_2 , 96p
Inception (5b)	$3 \times 3 \times 736$	256	96	384			m, 96p
Avg pool	736						
Linear (fc)	128						
l_2 normalisation	128						
Linear	N						
Binarisation	N						
Linear	128						
l_2 normalisation	128						

Note: Each row is a layer in the neural network and the last six columns indicate the parameters of pooling or the inception layers from [37]. This model is almost identical to the one described in [37]. The two major differences are the use of l_2 pooling instead of max pooling (m), where specified. That is, instead of taking the spatial max the l_2 norm is computed. The pooling is always 3×3 (aside from the final average pooling) and in parallel to the convolutional modules inside each Inception module. If there is a dimensionality reduction after the pooling it is denoted with p. 1×1 , 3×3 , and 5×5 pooling are then concatenated to get the final output.

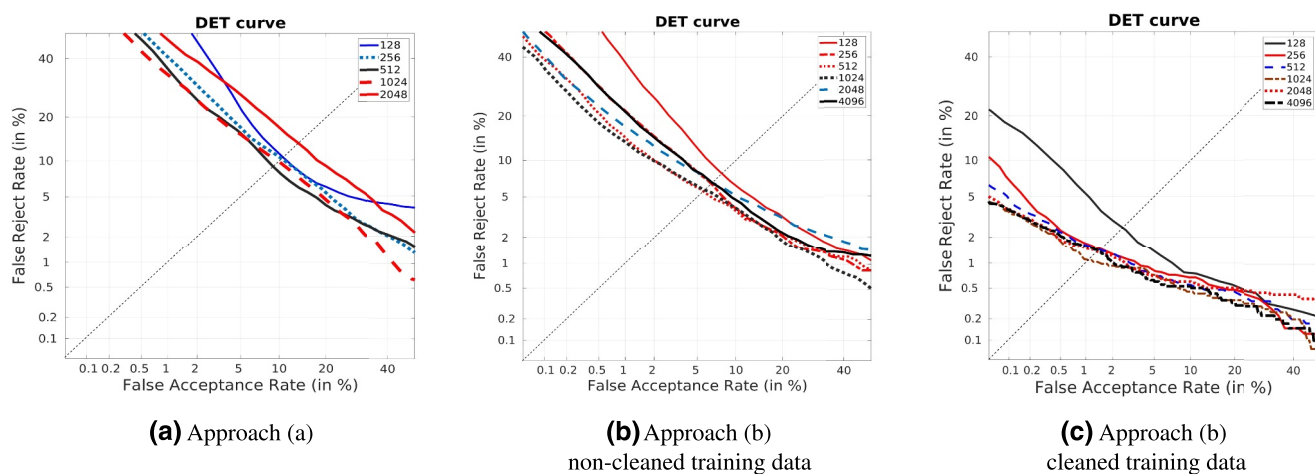
**FIGURE A1** DET curves of the Eval male partition of the MOBIO database using the standard protocol [32]. The training of the models is done using the MS-celeb-1M. Approach (a) denotes training from scratch, where approach (b) means training using pretrained models

TABLE A2 Performance on the MOBIO database using the standard protocol [32]. The performance metric is the half total error rate (HTER)

Length	Approach (a)		Approach (b) (non-cleaned training data)		Approach (b) (cleaned training data)	
	HTER (%) Eval female	HTER (%) Eval male	HTER (%) Eval female	HTER (%) Eval male	HTER (%) Eval females	HTER (%) Eval male
128	20.10	11.37	21.55	7.87	6.00	2.48
256	15.77	10.27	12.41	6.83	5.00	1.35
512	17.54	9.35	11.37	5.42	4.34	1.51
1024	16.60	14.09	9.82	5.48	5.26	1.27
2048	17.93	16.52	13.64	6.45	4.32	1.33
4096	46.80	47.26	25.87	10.68	4.29	1.38

Note: The training of the models is done using the MS-celeb-1M. Approach (a) denotes training from scratch, where approach (b) means training using pretrained models.