Deep Multi-biometric Fusion for Audio-Visual User Re-Identification and Verification

Mirko Marras[1], Pedro A. Marín-Reyes[2(✉)], Javier Lorenzo-Navarro[2], Modesto Castrillón-Santana[2], and Gianni Fenu[1]

[1] Department of Mathematics and Computer Science, University of Cagliari, V. Ospedale 72, 09124 Cagliari, Italy
{mirko.marras,fenu}@unica.it

[2] Instituto Universitario Sistemas Inteligentes y Aplicaciones Numericas en Ingenieria (SIANI), Universidad de Las Palmas de Gran Canaria, Campus Universitario de Tafira, 35017 Las Palmas de Gran Canaria, Spain
pedro.marin102@alu.ulpgc.es, {javier.lorenzo,modesto.castrillon}@ulpgc.es

Abstract. From border controls to personal devices, from online exam proctoring to human-robot interaction, biometric technologies are empowering individuals and organizations with convenient and secure authentication and identification services. However, most biometric systems leverage only a single modality, and may face challenges related to acquisition distance, environmental conditions, data quality, and computational resources. Combining evidence from multiple sources at a certain level (e.g., sensor, feature, score, or decision) of the recognition pipeline may mitigate some limitations of the common uni-biometric systems. Such a fusion has been rarely investigated at *intermediate level*, i.e., when uni-biometric model parameters are jointly optimized during training. In this chapter, we propose a multi-biometric model training strategy that digests face and voice traits in parallel, and we explore how it helps to improve recognition performance in re-identification and verification scenarios. To this end, we design a neural architecture for jointly embedding face and voice data, and we experiment with several training losses and audio-visual datasets. The idea is to exploit the relation between voice characteristics and facial morphology, so that face and voice uni-biometric models help each other to recognize people when trained jointly. Extensive experiments on four real-world datasets show that the biometric feature representation of a uni-biometric model jointly trained performs better than the one computed by the same uni-biometric model trained alone. Moreover, the recognition results are further improved by embedding face and voice data into a single shared representation of the two modalities. The proposed fusion strategy generalizes well on unseen and unheard users, and should be considered as a feasible solution that improves model performance. We expect that this chapter will support the biometric community to shape the research on deep audio-visual fusion in real-world contexts.

M. De Marsico et al. (Eds.): ICPRAM 2019, LNCS 11996, pp. 136–157, 2020.
https://doi.org/10.1007/978-3-030-40014-9_7

Keywords: Multi-biometric system · Cross-modal biometrics · Deep biometric fusion · Audio-visual learning · Verification · Re-identification

1 Introduction

Over the years, from visitors identified in human-robot interactions [27,28,46] to learners authenticated in online education platforms [13,14], biometrics has been increasingly playing a primary role in various contexts, such as robotics, medicine, science, engineering, education and several other business areas [25]. Evidence of this can be retrieved in recent reports that estimate a huge growth of the biometric market size, moving from $10.74 billion in 2015 to $32.73 billion by 2022[1]. Examples of biometric traits include the facial structure [49], the ridges of a fingerprint [35], the iris pattern [4], the sound waves of a voice [16], and the way a person interacts with a digital device [56]. From the system perspective, recognition pipelines detect the modality of interest in the biometric sample. This is followed by a set of pre-processing functions. Features are then extracted from pre-processed data, and used by a classifier for recognition. From the user perspective, an individual is asked to provide some samples whose feature vectors are stored as a template by the system (i.e., enrollment). Then, the recognition process may involve associating an identity with the probe (i.e., re-identification) or determining if the probe comes from the declared person (i.e., verification).

Most biometric systems manipulate a single modality (e.g., face only), and may encounter problems due to several factors surrounding the system and the user, such as the acquisition distance and the environmental conditions [3,10,44]. Deploying such systems in real-world scenarios thus presents various challenges. For instance, facial images exhibit large variations due to occlusions, pose, indoor-illumination, expressions, and accessories [49]. Similarly, audio samples vary due to the distance of the subject from the microphone, indoor reverberations, background noise, and so on [16]. Given the highly-variable conditions of these scenarios, focusing exclusively on one modality might seriously decrease the system reliability, especially when the acquisition conditions are not controlled. Multi-biometric systems have been proven to overcome some limitations of uni-biometric systems by combining evidence from different sources. This often results in improved recognition performance and enhanced system robustness, since the combined information is likely to be more distinctive compared to the one obtained from a single source [2,12]. Multi-biometric systems might be exploited in several scenarios, such as when people are speaking while being assisted by robots or when learners are attending an online oral exam.

One of the main design choices while developing a multi-biometric system is to select the level of the recognition pipeline where the fusion happens. To provide a unique global response, fusion policies generally refer to sensor level, feature level, score level, or decision level [41]. First, sensor-level fusion corresponds to combining raw data immediately after acquisition. Second, feature-level fusion refers to performing fusion of feature vectors extracted from different biometric

[1] https://www.grandviewresearch.com/industry-analysis/biometrics-industry.

samples. Third, score-level fusion corresponds to fuse matching scores produced by different systems. Fourth, decision-level fusion implies the combination of decisions taken by more than one system based on voting. Late fusion strategies usually made the process simpler and flexible, but an excessive amount of information entropy was lost. Early fusion policies were proven to work better, but tended to introduce high complexity and less flexibility. The recent revolution driven by deep-learned representations has contributed to reduce the latter deficiencies, and facilitated the experimentation of cost-effective intermediate fusion during model training and deployment [36]. It follows that, for instance, face and voice models might be trained jointly to learn whether face or voice probes come from a given user, but then deployed in a uni-biometric manner. On the other side, they could be combined in a multi-biometric way by embedding face and voice data into a single feature vector during deployment.

In this chapter, we introduce a multi-biometric training strategy that digests face and voice traits, and we investigate how it makes it possible to improve recognition performance in audio-visual re-identification and verification. With this in mind, we design a neural architecture composed by a sub-network for faces and a sub-network for voices fused at the top of the network and jointly trained. By exploiting features correlation, both models help each other to predict whether facial or vocal probes come from a given user. This paper extends the work presented in [30] that introduced an audio-visual dataset collected during human-robot interactions, and evaluated it and other challenging datasets on uni-biometric recognition tasks. The mission is to make a step forward towards the creation of biometric models able to work well on challenging real-world scenarios, such as identification performed by robots or continuous device authentication. More precisely, this paper provides the following contributions:

- We present a deeper contextualization of the state-of-the-art biometric solutions explored by researchers in audio-visual real-world biometrics scenarios.
- We experiment with a fusion strategy that combines face and voice traits instead of using them individually for re-identification and verification tasks.
- We extensively validate our strategy in public datasets, showing that it significantly improves uni-biometric and multi-biometric recognition accuracy.

Experiments on four datasets from real-world contexts show that the jointly-trained uni-biometric models reach significantly higher recognition accuracy than individually-trained uni-biometric models, both when their embeddings are deployed separately (i.e., 10%–20% of improvement) and when they are combined into a single multi-biometric embedding (i.e., 30%–50% of improvement). As the proposed strategy well generalizes on unseen and unheard users, it should be considered as a feasible solution for creating effective biometric models.

The rest of this chapter guides readers along the topic as follows. Section 2 summarizes recent uni-biometric strategies for face and voice recognition tasks together with biometric fusion strategies involving them. Then, Sect. 3 describes the proposed fusion strategy, including its formalization, input data formats, network architecture structures, and training process steps. Sections 4 depicts the

experimental evaluation of the proposed strategy, and highlights how it outperforms state-of-the-art solutions. Finally, Sect. 5 depicts conclusions, open challenges and future directions in this research area.

2 Related Work

In this section, we briefly describe state-of-the-art contributions on audio-visual biometrics applied in different scenarios. This is achieved by introducing face and voice uni-biometric systems and, subsequently, existing multi-biometric models.

2.1 Deep Face Recognition

The recent widespread of deep learning in different areas has favoured the usage of neural networks as feature extractors combined with common machine-learning classifiers, as proposed in [50]. Backbone architectures that accomplish this task rapidly evolved from *AlexNet* [24] to *SENet* [20] over last years.

In parallel, researchers have formulated both data sampling strategies and loss functions to be applied when such backbone architectures are trained. *Deepface* [44] integrates a cross-entropy-based *Softmax* loss while training the network. However, applying *Softmax* loss is usually not sufficient by itself to learn features separated by a large margin when the samples come from diverse entities, and other loss functions have been explored to enhance the generalization ability. For instance, euclidean-distance-based losses embed images into an euclidean space and reduce intra-variance while enlarging inter-variance across samples. *Contrastive* loss [43] and *Triplet* loss [38] are commonly used to this end, but they often exhibit training instability and complex sampling strategies. *Center* loss [51] and *Ring* loss [55] balance the trade-off between accuracy and flexibility. Furthermore, cosine-margin-based losses, such as *AM-Softmax*, were proposed to learn features separable through angular distance measures [48].

2.2 Deep Voice Recognition

Traditional speaker recognition systems based on hand-crafted solutions relied on *Gaussian Mixture Models* (GMMs) [37] that are trained on low dimensional feature vectors, *Joint Factor Analysis* (JFA) [11] methods that model speaker and channel subspaces separately, or *i-Vectors* [23] that attempt to embed both subspaces into a single compact, low-dimensional space.

Modern systems leveraged deep-learned acoustic representations, i.e., embeddings, extracted from one of the last layers of a neural network trained for standard or one-shot speaker classification [19,29]. The most prominent examples include *d-Vectors* [47], *c-Vectors* [7], *x-Vectors* [42], *VGGVox-Vectors* [34] and *ResNet-Vectors* [9]. Furthermore, deep learning frameworks with end-to-end loss functions to train speaker discriminative embeddings have recently drawn attention [18]. Their results proved that end-to-end systems with embeddings achieve better performance on short utterances common in several contexts (e.g., robotics, proctoring, and border controls) compared with hand-crafted systems.

2.3 Deep Audio-Visual Recognition

Combining signals from multiple sensors has been traditionally investigated from a data fusion perspective. For instance, such a merge step can happen at *sensor level* or *feature level*, and focuses on how to combine data from multiple sources, either by removing correlations between modalities or representing the fused data in a common subspace; the fused data is then fed into a machine-learning algorithm [1]. The literature provides also evidence of fusion techniques at *score level* and *decision level* [8,21,32,39]. There is no a general conclusion on which fusion policy performs better between early and late fusion, and the performance is problem-dependent [41]. However, late fusion was simpler to be implemented, particularly when modalities varied in dimensionality and sampling rates.

Emerging machine-learning strategies are making it possible to fill this gap in flexibility between early and late fusion. Through a new form of multi-biometric fusion of features representations, namely *intermediate fusion*, neural networks offer a flexible approach to multi-biometric fusion for numerous practical problems [36]. Given that neural architectures learn a hierarchical representation of the underlying data across its hidden layers, learned representations of different modalities can be fused at various levels of abstraction, introducing several advantages with respect to previous solutions [39,40]. Modality-wise and shared representations are learned from data, while features were originally manually designed and required prior knowledge on the data. Such a new fusion level requires little or no pre-processing of input data, differently from traditional techniques that may be sensitive to data pre-processing. Furthermore, implicit dimensionality reduction within the architecture and easily scalable capabilities are guaranteed, improving flexibility and accuracy at the same time.

Good evidence of these advantages comes from the literature. For instance, the authors in [15] aimed to learn features from audio and faces from convolutional neural networks compatible at high-level. Their strategy has been proven to produce better performance than single modality, showing the effectiveness of the multi-biometric fusion during deployment. The works in [5,6] proposed time-dependent audio-visual models adapted in an unsupervised fashion by exploiting the complementary of multiple modalities. Their approach allowed to control the model adaptation and to cope with situations when one of the two modalities is under-performing. Furthermore, the approach described in [45] used a three-dimensional convolutional neural network to map both modalities into a single representation space, and evaluated the correspondence of audio–visual streams using such learned multi-biometric features. Inspired by findings on high-level correlation of voice and face across humans, the authors in [40] experimented with an attention-based neural network that learns multi-sensory associations for user verification. The attention mechanism conditionally selects a salient modality representation between speech and facial ones, balancing between complementary inputs. Differently, the method in [52] extracted static and dynamic face and audio features; then, it concatenated the top discriminative visual-audio features to represent the two modalities, and used a linear classifier for identification. Recent experience in [26] depicted an efficient attention-guided audio-face

fusion approach to detect speakers. Their factorized model deeply fused the paired audio-face features, whereby the joint audio-face representation can be reliably obtained. Finally, the authors in [33] investigated face-voice embeddings enabling cross-modal retrieval from voice to face and vice versa.

The collection of large amounts of training data and the advent of powerful graphics processing units (GPUs) is enabling deep intermediate fusion, and this paper makes a step forward towards its application in the audio-visual domain.

3 The Proposed Intermediate Fusion Approach

In this section, we describe our intermediate fusion strategy that jointly learns voice and face embeddings, including model formalization, input data formats, underlying architectures, and training details (Fig. 1).

The core idea is to leverage the morphological relations existing between voice and face biometrics in order to investigate a cross-modal training where each uni-biometric model is supported by the biometric model of the other modality in improving the effectiveness of its feature representations. Differently from other intermediate fusion approaches, such a multi-biometric fusion might happen (i) on training to develop better uni-biometric models and/or (ii) on deployment to exploit joint evidence from the two modalities simultaneously.

Face Backbone Formalization. Let $A_f \subset \mathbb{R}^{m \times n \times 3}$ denote the domain of RGB images with $m \times n \times 3$ size. Each image $a_f \in A_f$ is pre-processed in order to detect the bounding box and key points (two eyes, nose and two mouth corners) of the face. The affine transformation is used to align the face. The image is then resized and each pixel value is normalised in the range [0, 1]. The resulting intermediate facial image, defined as $S_f \subset \mathbb{R}^{m \times n \times 3}$, is used as input of the visual modality branch of our model. In this branch, an explicit feature extraction which produces fixed-length representations in $D_f \subset \mathbb{R}^e$. We denote such a stage as $\mathcal{D}_{f_{\theta_f}} : A_f \rightarrow D_f$. Its output is referred to as face feature vector.

Voice Backbone Formalization. Let $A_v \subset \mathbb{R}^*$ denote the domain of waveforms digitally represented by an intermediate visual acoustic representation $S_v \subset \mathbb{R}^{k \times *}$, such as a spectrogram or a filter-bank. Each audio $a_v \in A_v$ is converted to single-channel. The spectrogram is then generated in a sliding window fashion using a Hamming window, generating an acoustic representation s_v that corresponds to the audio a_v. Mean and variance normalisation is performed on every frequency bin of the spectrum. The resulting representation is used as input of the acoustic modality branch of our model. In this branch, an explicit feature extraction which produces fixed-length representations in $D_v \subset \mathbb{R}^e$. We denote such a stage as $\mathcal{D}_{v_{\theta_v}} : S_v \rightarrow D_v$. Its output is named voice feature vector.

Fusion Backbone Formalization. Let $D^{2 \times e}$ be the domain of audio-visual feature vectors generated by a plain concatenation of the sparse representation from the face and voice backbones, i.e., d_f and d_v. We denote as $\mathcal{C}_\theta : (D_f, D_v) \rightarrow D^{2 \times e}$ such a concatenation stage of both modalities applied after the representation layer of each single modality branch. Then, an additional feature vector learning

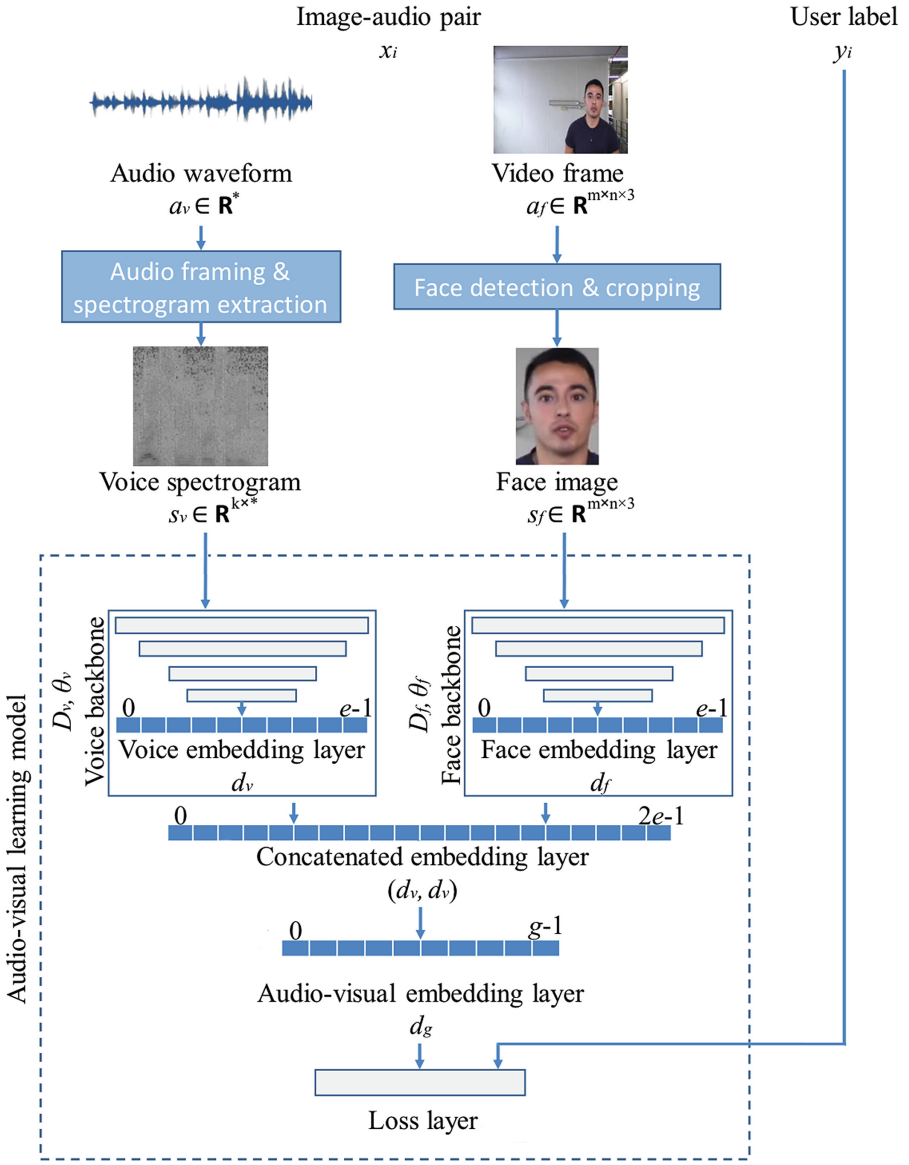

Fig. 1. The proposed neural architecture for intermediate multi-biometric fusion.

step is applied to the concatenated vector $d \in D^{2 \times e}$ to get a single feature vector of size g jointly learned from d_f and d_v. This extra layer aims to (i) keep independent the multi-biometric embedding size from the uni-biometric embedding sizes and (ii) learn more compacted and flexible representations. Moreover, by setting $g = e$, reasonable comparisons between uni-biometric and multi-biometric sparse

representations of the same size can be performed. We denote such an extra step as $\mathcal{D}_{fv_{\theta_{f,v}}} : D^{2\times e} \rightarrow D^{g}$. Its output is named as audio-visual feature vector.

Combining both modalities might generate a better sparse representation of the individual, and enrich the feature representation of a single modality. This is due to the relations of voice to genre and facial morphology of people, e.g., male people commonly have a tone lower than female people. Therefore, by leveraging the fusion backbone, the uni-biometric backbones help each other to better recognize people. Our hypothesis is that the embeddings of each backbone should perform better when trained jointly than when trained separately.

Backbones Instantiation. The proposed approach makes use of existing neural network architectures, slightly arranged to accommodate the modality digested by each of the above-mentioned backbones and the subsequent fusion purposes.

Two instances of the residual-network (*ResNet-50*) architecture are used as feature vector extractors $\mathcal{D}_{f_{\theta_f}}$ and $\mathcal{D}_{v_{\theta_v}}$ within face and voice backbones, respectively [17]. Such a network, well known for good classification performance on visual and acoustic modalities [9,44], is similar to a multi-layer convolutional neural network, but with added skip connections such that the layers add residuals to an identity mapping on the channel outputs. The input layers of the original *ResNet-50* architecture are adapted to the modality associated to the corresponding each backbone. Moreover, the fully-connected layer at the top of the original network is replaced by two layers: a flatten layer and a fully-connected layer whose output is the embedding of the modality, i.e., d_f or d_v.

The fusion backbone $\mathcal{D}_{fv_{\theta_{f,v}}}$ is instantiated by a concatenation layer stacked into the model to combine face and voice feature vectors in $D^{2\times e}$ domain, and an additional fully-connected layer where the significant features of video and audio modality are jointly embedded. The latter output represents the audio-visual feature vector $d \in D^{g}$ previously formalized. Moreover, for each fully-connected layer, batch normalization has been set before the activation function to regularize the outputs, and a dropout layer is inserted after activation to prevent model over-fitting. Finally, an output layer depending on the applied loss function is posed at the top of the network during training.

Training Process Description. The training data is composed by N tuples $\{(x_i, y_i)\}_{i=1}^{N}$ where each multi-biometric sample x_i corresponds to a person associated with the class $y_i \in 1, ..., I$, being I the number of different identities depicted in N samples. Each sample x_i is defined as a pair $x_i = (a_{v_i}, a_{f_i})$ such that a_{v_i} is a utterance and a_{f_i} is a visual frame. The elements of each pair are randomly chosen among face and voice samples from the same user; then, they are sequentially fed into the multi-biometric model. Such a model can be integrated with any existing loss function. Additionally, a hold-out validation set consisting of all the speech and face segments from a single randomly-selected video per user is used to monitor training performance.

4 Experimental Evaluation

In this section, we assess the effectiveness of our fusion strategy. First, we detail the datasets, the experimental protocols, the implementation details, and the loss functions. Then, we present the results achieved by the fusion strategy on re-identification and verification, varying the loss function and the testing dataset.

4.1 Training and Testing Datasets

We considered traditional audio-visual datasets for training the models, and we tested them on datasets from diverse audio-visual contexts (see Fig. 2). This

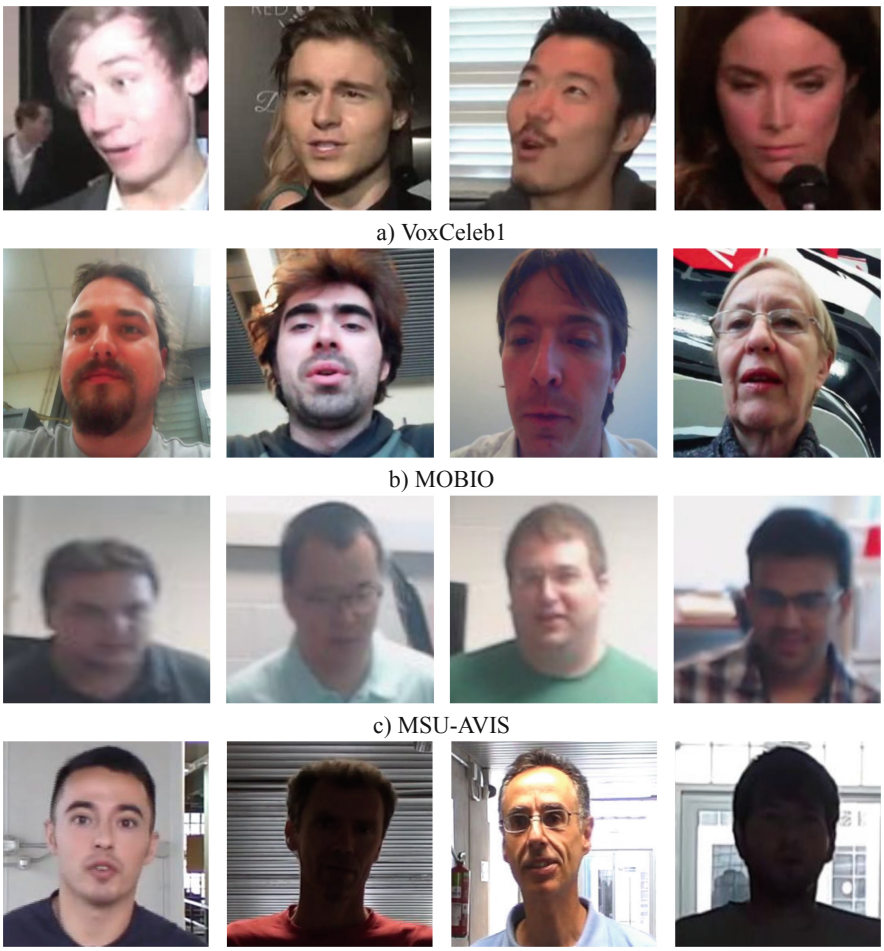

Fig. 2. Facial samples coming from the testing datasets used to evaluate our approach.

choice enables the computation of additional state-of-the-art benchmark scores on *AveRobot*, and make it possible to observe how the strategy affects the performance on different contexts. The audio-visual datasets are divided in one training dataset and four testing datasets to replicate a cross-dataset setup:

- **Training Dataset.** *VoxCeleb1-Dev* is an audio-visual speaker identification and verification dataset collected by [34] from Youtube, including 21,819 videos from 1,211 identities. It is the one of the most suited for training a deep neural network due to the wide range of users and samples per user.
- **Testing Dataset #1.** *VoxCeleb1-Test* is an audio-visual speaker identification and verification dataset collected by [34] from Youtube, embracing 677 videos from 40 identities.
- **Testing Dataset #2.** *MOBIO* is a face and speaker recognition dataset collected by [31] from laptops and mobile phones under a controlled scenario, including 28,800 videos from 150 identities.
- **Testing Dataset #3.** *MSU-Avis* is a face and voice recognition dataset collected by [8] under semi-controlled indoor surveillance scenarios, including 2,260 videos from 50 identities.
- **Testing Dataset #4.** *AveRobot* is an audio-visual biometric recognition dataset collected under robot assistance scenarios in [30], including 2,664 videos from 111 identities.

The reader notices that acquisition distance, environmental conditions, and data quality greatly vary among the datasets, making them challenging.

4.2 Evaluation Setup and Protocols

Experiments aimed to assess both uni-biometric and multi-biometric feature representations through evaluation protocols applied in re-identification and verification tasks (Fig. 3).

Tested Data Format. For the face branch, each frame is analyzed in order to detect the face area and landmarks through MTCNN [53]. The five facial points (two eyes, nose and two mouth corners) are adopted by such an algorithm to perform face alignment. The faces are then resized to 112×112 pixels in order to fit in our branch and each pixel in *[0, 255]* in RGB images is normalized by subtracting *127.5* then dividing by *128*. The resulting images are then used as input to the face branch. For the voice branch, each audio is converted to single-channel, *16*-bit streams at a *16* kHz sampling rate for consistency. The spectrograms are then generated in a sliding window fashion using a Hamming window of width 25 ms and step 10 ms. This gives spectrograms of size 512×300 for three seconds of speech. Mean and variance normalisation is performed on every frequency bin of the spectrum. No other speech-specific pre-processing is used. The spectrograms are used as input to the voice branch.

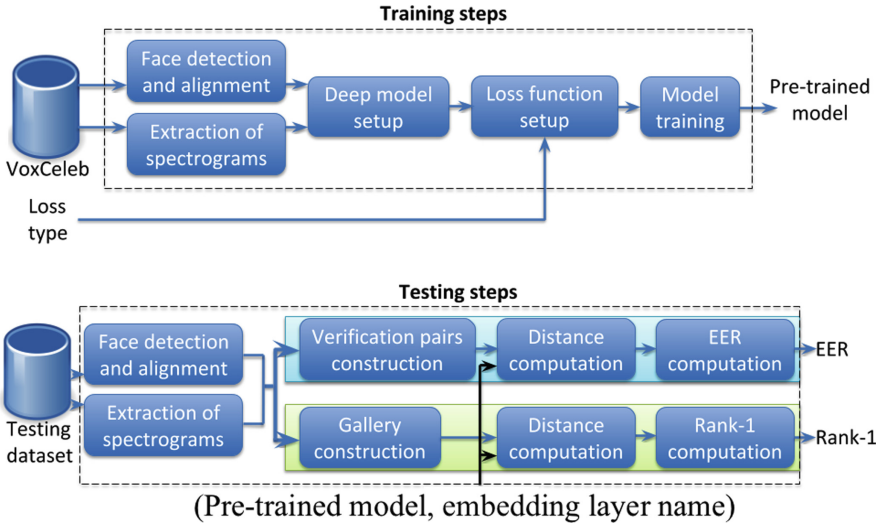

Fig. 3. Experimental evaluation overview. Training and testing protocols.

Tested Feature Representations. The evaluation involved uni-biometric and multi-biometric feature representations obtained from backbone networks trained on top of *VoxCeleb1-Dev*. In order to optimize model weights, several instances of the network were independently trained through different loss functions from various families: *Softmax* loss [44], *Center* loss [51], *Ring* loss [55], and *AM-Softmax* loss [48]. More precisely, for each training loss, we trained appropriate models to learn the following feature representations:

- *Uni-Modal Voice* representations extracted from d_v when the voice branch is trained alone (baseline).
- *Uni-Modal Face* representations extracted from d_f when the face branch is trained alone (baseline).
- *Multi-Modal Voice* representations extracted from d_v when the voice branch is trained jointly with the face branch (introduced in this paper).
- *Multi-Modal Face* representations extracted from d_f when the face branch is trained jointly with the voice branch (introduced in this paper).
- *Multi-Modal Face+Voice* representations extracted from d_g when the face branch and the voice branch are jointly trained (introduced in this paper).

Each model was initialised with weights pre-trained on ImageNet. Stochastic gradient descent with a weight decay set to *0.0005* was used on mini-batches of size *512* along *40* epochs. The initial learning rate was *0.1*, and this was decreased with a factor of *10* after *20*, *30* and *35* epochs. The training procedure was coded in Python, using Keras on top of Tensorflow.

Re-identification Protocol. For each testing dataset, the protocol aims to evaluate how the learned representation are capable of predicting, for a given

test frame/spectrogram, the identity of the person chosen from a gallery of identities. For each experiment conducted on a testing dataset, we randomly selected 40 users every time in order to (i) keep constant the number of considered users, and (ii) maintain comparable the results across the different datasets. *VoxCeleb1-Test* has the minimum number of participants among the considered datasets (i.e., *40*). For each user, we have chosen the first *80%* of videos for the gallery, while the other *20%* of videos were probes. For each user, we randomly selected *20* frames/spectrograms from the gallery videos as gallery images, and *100* frames/spectrograms from the probe videos as probe images. Then, given each frame/spectrogram, the corresponding feature representation was extracted. The *Euclidean* distance was used to compare feature vectors obtained from models trained on *Softmax*, *Center* loss and *Ring* loss, while the *Cosine* distance was used for features vectors obtained from models trained on *AM-Softmax* loss due to its underlying design. Then, we measured the top one rank, a well-accepted measure to evaluate the performance on people re-identification tasks (e.g., [54]). The probe image is matched against a set of gallery images, obtaining a ranked list according to their matching similarity/distance. The correct match is assigned to one of the top ranks, the top one rank in this case (*Rank-1*).

The *Rank-1* is formulated as the accuracy on predicting the right identity (prediction) given the known spectrogram/face identity (ground truth):

$$Rank\text{-}1 = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP is the true positive, TN represents the true negative, FP is the false positive and FN represents the false negatives. Thus, it was used to evaluate the performance of the models on the test images/spectrograms. Starting from the subject selection, the experiment was repeated and the results were averaged.

Verification Protocol. For each testing dataset, the protocol aims to evaluate how the learned representations are capable of verifying, given a pair of test frames/spectrograms, whether the faces/voices come from the same person. From each testing dataset, we randomly selected *40* subjects due to the same reasons stated in the above re-identification protocol. Then, we randomly created a list of *20* videos (with repetitions) for each selected user and, from each one of them, we randomly created *20* positive frame pairs and *20* negative frame pairs. The above-mentioned feature representations were considered as feature vector associated to each frame/spectrogram. We used the same distance measures leveraged for re-identification and the *Equal Error Rate* (EER) was computed to evaluate the performance of the models on the test pairs. *EER* is a well-known biometric security metric measured on verification tasks [22]. *EER* indicates that the proportion of false acceptances (*FAR*) is equal to the proportion of false rejections (*FRR*). Both measures are formulated as:

$$\begin{aligned} FAR &= \frac{\text{number of false accepts}}{\text{number of impostors comparisons}} \\ FRR &= \frac{\text{number of false rejects}}{\text{number of genuine comparisons}} \end{aligned} \tag{2}$$

The lower the EER, the higher the performance. Lastly, starting from the subject selection, the experiment was repeated and the results were averaged.

4.3 Re-Identification Results

The *Rank-1* performance on the testing datasets is shown in Figs. 4, 5, 6 and 7. It can be observed that the results vary with respect to the modality, training loss, and the dataset. Results are presented from these three point of views.

Considering the face modality, the representations learned through the *Softmax* loss appear as the best performer for the uni-modal face setup (*Rank-1* from *0.32* to *0.74*), while the representations performance for the multi-modal face setup greatly varies among the training losses and the testing datasets. This means that the deep multi-biometric training strategy is strongly affected by the training loss and the targeting dataset, while common uni-biometric training strategies take advantage of the *Softmax* loss and their results are affected only by the dataset. Furthermore, it can be observed that multi-modal face representations make it possible to improve the results in face re-identification on challenging scenarios as in the *AveRobot* dataset. In more controlled scenarios, while the deep fusion allows us to increase the accuracy, it reaches results comparable with the ones obtained by uni-modal features representations learned through *Softmax* loss in uni-biometric face models.

Different observations can be made for the voice modality. The representations learned through the *Center* loss are superior to representations learned by other losses in uni-modal and multi-modal voice models. Interestingly, the multi-modal voice representations perform worse than the uni-modal voice representations for any loss. It follows that voice biometrics does not take a large advantage of the deep fusion strategy, differently from what happens for face biometrics. The exception is represented by results obtained in *MOBIO* multi-modal voice representations; they reach higher results than the uni-modal voice representations. This means that there is not a general conclusion regarding the effectiveness of multi-modal voice representations, but they are dataset-dependent. Therefore, preliminary tests should be performed to select the right voice training strategy based on the context.

In the case both face and voice biometrics are fused (*multi-modal face+voice* setup), the representations learned through *Ring* and *Center* losses achieved better results than uni-modal representations, while the representations learned through *Softmax* and *AM-Softmax* losses reach worse results probably due to the bad performance of the intermediate multi-modal voice representation. It follows that a plain fusion of face and voice embeddings during training is not always sufficient to improve the results with respect to uni-modal representations. It appears necessary to design countermeasures for controlling the contribution of each modality on the audio-visual embedding during training.

Among the datasets, *VoxCeleb1-Test* showed the highest *Rank-1* values. This is probably related to the fact that all the models are trained on data coming from the same context of the testing dataset. On *MOBIO*, the representations tended to achieve comparable results as the data includes front-face videos recorded

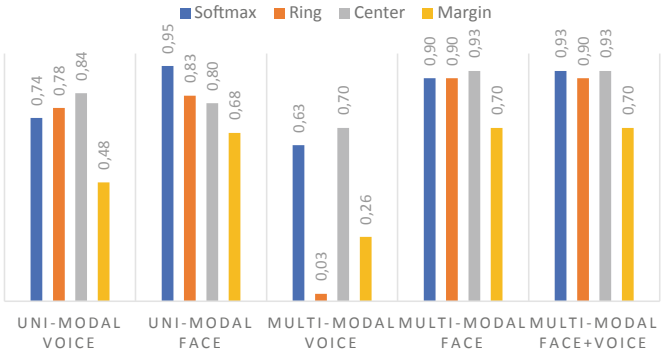

Fig. 4. Re-identification results on VoxCeleb1-Test - Rank-1.

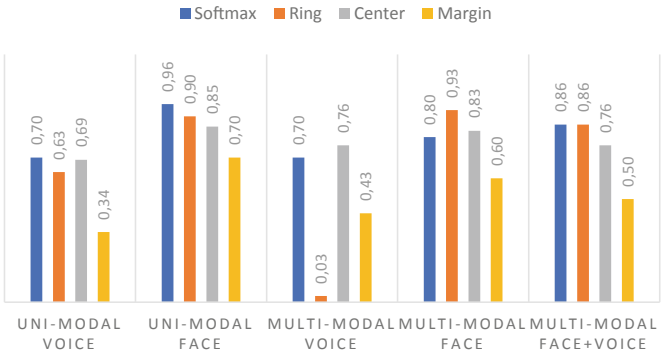

Fig. 5. Re-identification results on MOBIO - Rank-1.

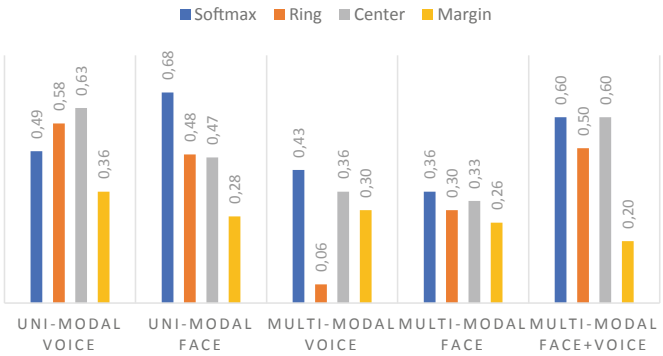

Fig. 6. Re-identification results on MSU-Avis - Rank-1.

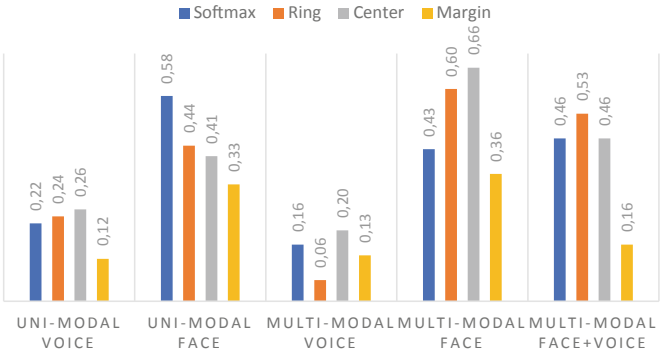

Fig. 7. Re-identification results on Averobot - Rank-1.

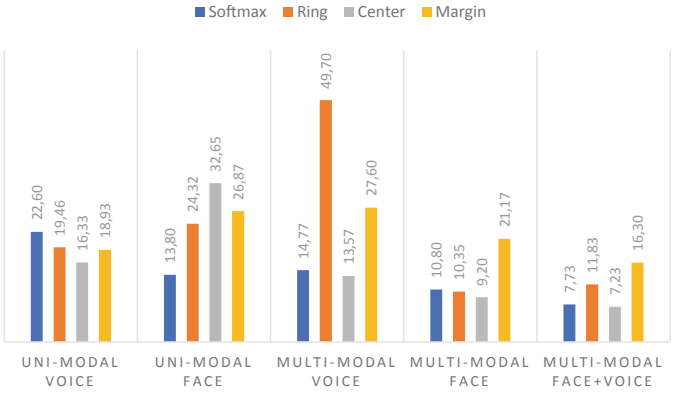

Fig. 8. Verification results on VoxCeleb1-Test - EER.

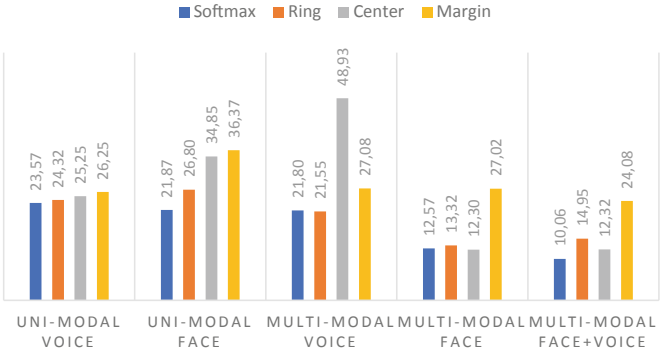

Fig. 9. Verification results on MOBIO - EER.

from the smartphone, i.e., a controlled conditions where the recognition should be easier. Differently, *MSU-Avis* and *AveRobot* highlight several challenges for the trained representations. The more uncontrolled scenarios conveyed by the latter datasets are the main reasons of the significantly lower *Rank-1* values. In particular, the *AveRobot* dataset represents the most challenging scenario, and more effective fusion strategies should be designed starting from the one presented in this paper.

4.4 Verification Results

Figures 8, 9, 10 and 11 plot the results achieved by the learned representations on verification. The ranking is slightly different with respect to the re-identification task, and the impact of the context, the loss, and the modality varies across settings.

It can be observed that multi-modal face representations achieve lower *EER* than uni-modal face representations with all the dataset and training losses. This means that deep fusion significantly helps to create better sparse representations for the face modality. More precisely, *EER* obtained by representations learned through *Ring* and *Center* losses can be improved of around *50%*, while we observed an improvement of around *25%* thanks to representations learned through *Softmax* and *Margin* losses. It follows that multi-modal face sparse representations better separate among genuine and impostor pairs.

Comparable results are obtained by multi-modal voice representations, even though the improvement with respect to the uni-modal voice representations is less evident, i.e. among *5%* and *10%*. Interestingly, multi-modal voice representations learned through *Ring* loss do not work well. It follows that, as shown on the re-identification task, the *Ring* loss suffers from the deep fusion approach. Our results suggest that such a loss has a minor impact in deep audio-visual fusion settings.

By merging face and voice embeddings into a single representation, the verification performance improves on all the datasets, with all the training losses. It can be observed an improvement of around *50%* on all the settings. The face-voice fused representations work well also when learned through *Ring* loss; hence, the deficiencies experienced by multi-modal voice representations learned through *Ring* loss are mitigated by fusing voices and faces.

The results across the testing datasets confirm the observations made for the re-identification task. The context has a relevant impact on the absolute performance of the models, moving from *VoxCeleb1-Test* to *AveRobot* by increasing challenging level. In particular, the verification results on *AveRobot* pairs are *4* or *5* times worse than the ones achieved on *VoxCeleb1-Test* pairs. The reasons behind this large difference could be related to the more uncontrolled conditions characterized by very dark faces and highly-noisy surroundings.

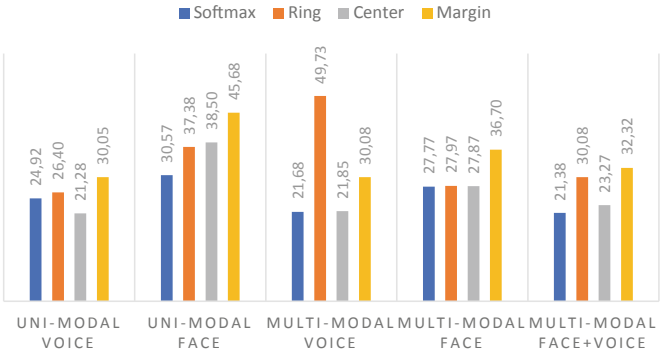

Fig. 10. Verification results on MSU-Avis - EER.

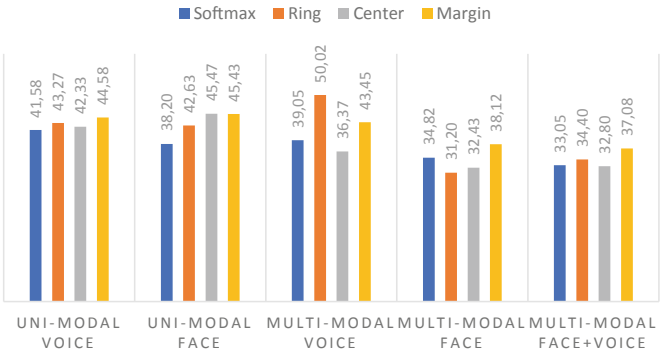

Fig. 11. Verification results on Averobot - EER.

5 Conclusions, Open Challenges, and Future Directions

In this chapter, we proposed a deep intermediate fusion strategy of audio-visual biometric data. By combining state-of-the-art deep learning methodologies, a two-branch neural network fed with face and voice pairs aimed to jointly learn uni-biometric and multi-biometric fixed-length feature representations by exploiting feature correlation. Branches influence each other in computing the right classification label after their fusion during training, so that the representation layer of each uni-biometric model performs better than the one returned by a uni-biometric model trained alone. The results were further improved by jointly learning a single audio-visual embedding that includes information from both face and voice evidence. Based on the obtained results, we can conclude that:

- Face and voice models can benefit from deep intermediate fusion, and the recognition improvement depends on the modality, the loss, and the context.
- Deep intermediate fusion during training can be used to significantly increase recognition accuracy of uni-biometric face and voice models.

- Uni-biometric face models exhibit higher accuracy improvements than uni-biometric voice models after being jointly trained at intermediate level.
- Merging face and voice into a single embedding vector at intermediate level positively impacts the accuracy of multi-biometric audio-visual models.
- Face and voice models jointly trained at intermediate level generalize well across populations and are more robust when applied in challenging contexts.
- Deep intermediate fusion should be considered as a viable solution for creating more robust and reliable biometric models.

Research on deep multi-biometric fusion has produced a variety of solid methods, but still poses some interesting challenges that require further investigation:

- **Deep Fusion Capability.** Techniques in deep multi-modal learning facilitate a flexible intermediate-fusion approach, which not only makes it simpler to fuse modality-wise representations and learn a joint representation but also allows multi-modal fusion at various depths in the architecture. Moreover, deep learning architectures still involve a great deal of manual design, and experimenters may not have explored the full space of possible fusion architectures. It is natural that researchers should extend the notion of learning to architectures adaptable to a specific task.
- **Transferability across Contexts.** Existing models tend to be sensitive to the context targeted by the underlying training data. This has favored the creation of biometric models that, after being trained with data from a given context, do not generalize well in other contexts. With the new availability of public datasets and pre-trained models, it will become easier to plug them into a task different from the original one. Researchers could fine-tune pre-trained models with small amounts of context-specific data.
- **Robustness in Uncontrolled Environments.** Devising audio-visual biometric systems that can operate in unconstrained sensing environments is another unsolved problem. Most biometric systems either implicitly or explicitly impose some constraints on the data acquisition. Such constraints have to be reduced in order to seamlessly recognize individuals, i.e., the interaction between an individual and a biometric system should be transparent. This necessitates innovative interfaces and robust data processing algorithms.
- **Robustness against Spoofing Attacks.** Synthetically generated traits or maliciously modified traits are used to circumvent biometric systems. The challenge is to develop counter-measures that are applicable to hither to unseen or unknown attacks. Evaluating and assessing how the deployment of multi-biometric systems might help to face this challenge requires further investigation.
- **Explainablity and Interpretability.** Most machine-learning algorithms built into automation and artificial intelligence systems lack transparency, and may contain an imprint of the unconscious biases of the data and algorithms underlying them. Hence, it becomes important to understand how we can predict what is going to be predicted, given a change in input or algorithmic parameters. Moreover, it requires attention how the internal mechanics of the system can be explained in human terms.

– **Fairness, Transparency and Accountability.** With the advent of machine-learning, addressing bias within biometric systems will be a core priority due to several reasons. For instance, some biases can be introduced by using training data which is not an accurate sample of the target population or is influenced by socio-cultural stereotypes. Moreover, the methods used to collect or measure data and the algorithms leveraged for predicting identities can propagate biases. Future research should control these biases in the developed models, promoting fair, transparent, and accountable systems.

We expect that the case study on audio-visual fusion covered in this chapter will help researchers and developers to shape future research in the field.

Acknowledgments. Mirko Marras gratefully acknowledges Sardinia Regional Government for the financial support of his PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2014–2020, Axis III "Education and Training", Thematic Goal 10, Priority of Investment 10ii, Specific Goal 10.5).

This research work has been partially supported by the Spanish Ministry of Economy and Competitiveness (TIN2015-64395-R MINECO/FEDER) and the Spanish Ministry of Science, Innovation and Universities (RTI2018-093337-B-I00), by the Office of Economy, Industry, Commerce and Knowledge of the Canary Islands Government (CEI2018-4), and the Computer Science Department at the Universidad de Las Palmas de Gran Canaria.

References

1. Abozaid, A., Haggag, A., Kasban, H., Eltokhy, M.: Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion. Multimed. Tools Appl. **78**, 1–17 (2018)
2. Barra, S., Casanova, A., Fraschini, M., Nappi, M.: Fusion of physiological measures for multimodal biometric systems. Multimed. Tools Appl. **76**(4), 4835–4847 (2017)
3. Barra, S., De Marsico, M., Galdi, C., Riccio, D., Wechsler, H.: FAME: face authentication for mobile encounter. In: 2013 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS), pp. 1–7. IEEE (2013)
4. Bowyer, K.W., Burge, M.J.: Handbook of Iris Recognition. Springer, Heidelberg (2016). https://doi.org/10.1007/978-1-4471-6784-6
5. Brutti, A., Cavallaro, A.: Online cross-modal adaptation for audio-visual person identification with wearable cameras. IEEE Trans. Hum.-Mach. Syst. **47**(1), 40–51 (2016)
6. Cavallaro, A., Brutti, A.: Audio-visual learning for body-worn cameras. In: Multimodal Behavior Analysis in the Wild, pp. 103–119. Elsevier (2019)
7. Chen, Y.H., Lopez-Moreno, I., Sainath, T.N., Visontai, M., Alvarez, R., Parada, C.: Locally-connected and convolutional neural networks for small footprint speaker recognition. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
8. Chowdhury, A., Atoum, Y., Tran, L., Liu, X., Ross, A.: MSU-AVIS dataset: fusing face and voice modalities for biometric recognition in indoor surveillance videos. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3567–3573. IEEE (2018)

9. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
10. Cruz, C., Sucar, L.E., Morales, E.F.: Real-time face recognition for human-robot interaction. In: 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008, pp. 1–6. IEEE (2008)
11. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011)
12. Fenu, G., Marras, M.: Leveraging continuous multi-modal authentication for access control in mobile cloud environments. In: Battiato, S., Farinella, G.M., Leo, M., Gallo, G. (eds.) ICIAP 2017. LNCS, vol. 10590, pp. 331–342. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70742-6_31
13. Fenu, G., Marras, M.: Controlling user access to cloud-connected mobile applications by means of biometrics. IEEE Cloud Comput. **5**(4), 47–57 (2018)
14. Fenu, G., Marras, M., Boratto, L.: A multi-biometric system for continuous student authentication in e-learning platforms. Pattern Recogn. Lett. **113**, 83–92 (2018)
15. Geng, J., Liu, X., Cheung, Y.M.: Audio-visual speaker recognition via multi-modal correlated neural networks. In: 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW), pp. 123–128. IEEE (2016)
16. Hansen, J.H., Hasan, T.: Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process. Mag. **32**(6), 74–99 (2015)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
18. Heigold, G., Moreno, I., Bengio, S., Shazeer, N.: End-to-end text-dependent speaker verification. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5115–5119. IEEE (2016)
19. Hershey, S., et al.: CNN architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131–135. IEEE (2017)
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 7 (2017)
21. Huang, L., Yu, C., Cao, X.: Bimodal biometric person recognition by score fusion. In: 2018 5th International Conference on Information Science and Control Engineering (ICISCE), pp. 1093–1097. IEEE (2018)
22. Jain, A., Hong, L., Pankanti, S.: Biometric identification. Commun. ACM **43**(2), 90–98 (2000)
23. Kanagasundaram, A., Vogt, R., Dean, D.B., Sridharan, S., Mason, M.W.: I-vector based speaker recognition on short utterances. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association, pp. 2341–2344. International Speech Communication Association (ISCA) (2011)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
25. Li, S.Z., Jain, A.: Encyclopedia of Biometrics. Springer, Heidelberg (2015)
26. Liu, X., Geng, J., Ling, H., Cheung, Y.M.: Attention guided deep audio-face fusion for efficient speaker naming. Pattern Recogn. **88**, 557–568 (2019)
27. López, J., Pérez, D., Santos, M., Cacho, M.: GuideBot. A tour guide system based on mobile robots. Int. J. Adv. Robot. Syst. **10**, 381 (2013)
28. López, J., Pérez, D., Zalama, E., Gomez-Garcia-Bermejo, J.: BellBot - a hotel assistant system using mobile robots. Int. J. Adv. Robot. Syst. **10**, 40 (2013)

29. Lukic, Y., Vogt, C., Dürr, O., Stadelmann, T.: Speaker identification and clustering using convolutional neural networks. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), 13–16 September 2016, Vietri sul Mare, Italy. IEEE (2016)
30. Marras, M., Marín-Reyes, P.A., Lorenzo-Navarro, J., Castrillón-Santana, M., Fenu, G.: AveRobot: an audio-visual dataset for people re-identification and verification in human-robot interaction. In: International Conference on Pattern Recognition Applications and Methods (2019)
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 (ICMEW), pp. 635–640. IEEE (2012)
32. Memon, Q., AlKassim, Z., AlHassan, E., Omer, M., Alsiddig, M.: Audio-visual biometric authentication for secured access into personal devices. In: Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science, pp. 85–89. ACM (2017)
33. Nagrani, A., Albanie, S., Zisserman, A.: Learnable PINs: cross-modal embeddings for person identity. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 71–88 (2018)
34. Nagrani, A., Chung, J.S., Zisserman, A.: VoxCeleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612 (2017)
35. Peralta, D., et al.: A survey on fingerprint minutiae-based local matching for verification and identification: taxonomy and experimental evaluation. Inf. Sci. **315**, 67–87 (2015)
36. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process. Mag. **34**(6), 96–108 (2017)
37. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digit. Signal Proc. **10**(1–3), 19–41 (2000)
38. 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)
39. Sell, G., Duh, K., Snyder, D., Etter, D., Garcia-Romero, D.: Audio-visual person recognition in multimedia data from the IARPA Janus program. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3031–3035. IEEE (2018)
40. Shon, S., Oh, T.H., Glass, J.: Noise-tolerant audio-visual online person verification using an attention-based neural network fusion. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3995–3999. IEEE (2019)
41. Singh, M., Singh, R., Ross, A.: A comprehensive overview of biometric fusion. Inf. Fusion **52**, 187–205 (2019)
42. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: robust DNN embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5329–5333. IEEE (2018)
43. Sun, Y., Liang, D., Wang, X., Tang, X.: DeepID3: face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
44. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)

45. Torfi, A., Iranmanesh, S.M., Nasrabadi, N., Dawson, J.: 3D convolutional neural networks for cross audio-visual matching recognition. IEEE Access **5**, 22081–22091 (2017)
46. Troniak, D., et al.: Charlie rides the elevator-integrating vision, navigation and manipulation towards multi-floor robot locomotion. In: 2013 International Conference on Computer and Robot Vision (CRV), pp. 1–8. IEEE (2013)
47. Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J.: Deep neural networks for small footprint text-dependent speaker verification. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4052–4056. IEEE (2014)
48. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Process. Lett. **25**(7), 926–930 (2018)
49. Wang, M., Deng, W.: Deep face recognition: a survey. arXiv preprint arXiv:1804.06655 (2018)
50. Wang, Y., Shen, J., Petridis, S., Pantic, M.: A real-time and unsupervised face re-identification system for human-robot interaction. Pattern Recogn. Lett. **128**, 559–568 (2018)
51. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31
52. Zhang, J., Richmond, K., Fisher, R.B.: Dual-modality talking-metrics: 3D visual-audio integrated behaviometric cues from speakers. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3144–3149. IEEE (2018)
53. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016). https://doi.org/10.1109/LSP.2016.2603342
54. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3), 653–668 (2013)
55. Zheng, Y., Pal, D.K., Savvides, M.: Ring loss: convex feature normalization for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5089–5097 (2018)
56. Zhong, Y., Deng, Y.: A survey on keystroke dynamics biometrics: approaches, advances, and evaluations. In: Recent Advances in User Authentication Using Keystroke Dynamics Biometrics, pp. 1–22 (2015)