Pattern Recognition Letters 149 (2021) 179–184

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Improving user verification in human-robot interaction from audio or image inputs through sample quality assessment

David Freire-Obregón, Kevin Rosales-Santana, Pedro A. Marín-Reyes, Adrian Penate-Sanchez*, Javier Lorenzo-Navarro, Modesto Castrillón-Santana

SIANI, Universidad de Las Palmas de Gran Canaria, Spain

ARTICLE INFO

Article history:
Received 20 July 2020
Revised 26 March 2021
Accepted 23 June 2021
Available online 3 July 2021

Edited by : Maria De Marsico

MSC:
41A05
41A10
65D05
65D17

Keywords:
Biometric verification
Audiovisual verification
Human robot interaction

ABSTRACT

In this paper, we tackle the task of improving biometric verification in the context of Human-Robot Interaction (HRI). A robot that wants to identify a specific person to provide a service can do so by either image verification or, if light conditions are not favourable, through voice verification. In our approach, we will take advantage of the possibility a robot has of recovering further data until it is sure of the identity of the person. The key contribution is that we select from both image and audio signals the parts that are of higher confidence. For images we use a system that looks at the face of each person and selects frames in which the confidence is high while keeping those frames separate in time to avoid using very similar facial appearance. For audio our approach tries to find the parts of the signal that contain a person talking, avoiding those in which noise is present by segmenting the signal. Once the parts of interest are found, each input is described with an independent deep learning architecture that obtains a descriptor for each kind of input (face/voice). We also present in this paper fusion methods that improve performance by combining the features from both face and voice, results to validate this are shown for each independent input and for the fusion methods.

1. Introduction

Robot interaction with humans in many cases requires for the robot to be able to remember the identity of the individual with whom it is interacting. This is necessary when guiding people in public buildings like museums. Particularly in such public buildings, it is frequent that a team of robots cooperate providing services in different floors, simplifying the robot navigation across different floors, as they are not needed to manage stairs and lifts which introduces additional complexity [22,23,36]. In such situations user data, mainly non-cooperative biometric information, is shared among robots. Thus, the face and the voice together with other soft biometrics descriptors may be used to re-identify or verify the casual or anonymous (as he/she is not previously registered) user identity [3].

In this paper, we tackle identity verification from the different inputs a robot has (audio and video). We will show that by understanding the context and taking advantage of the fact that a robot can look or hear twice substantial improvements can be made. We provide solutions that improve on the state of the art by carefully analysing what parts of the input signal are relevant to perform biometric verification. Results for all three possible scenarios are provided: only face, only voice and a fusion of face + voice. As it can be expected, and proven in the experimental section of this paper, the best approach is the one that combines face biometric verification and voice biometric verification to perform user verification.

2. Related work

In the context of human-robot interaction (HRI), service robots play a main role with applications in a wide set of scenarios. Robot Minerva was one of the pioneer tour guiding robots [35] performing in the Smithsonian's National Museum of American History in Washington. Other robots appeared later in different museums across the world exhibiting similar capabilities: Robovie at the Osaka Science Museum [31], KTBot at the Eureka Science Museum of San Sebastian [34], and Robotinho at the Deutsches Museum of Bonn [8]. Mostly, robots actions are limited to a single floor. Considering the multirobot-human interaction in different buildings,

* Corresponding author .
E-mail address: adrian.penate@ulpgc.es (A. Penate-Sanchez).

https://doi.org/10.1016/j.patrec.2021.06.014

the robot system network proposed in [10] guide customers in a shopping mall, integrating sensors and cloud resources. More recently, the authors of [28] describe GidaBot, a heterogeneous service multi-robot team which cooperate in different floors. Later, they integrated face recognition to verify user identity when receiving a user in a different floor, providing real life results for 56 guiding actions, i.e. identities [29].

Face and audio are the main cues used to recognize individuals during HRI and feature fusion has also been analyzed in the HRI scenario. In [26] the authors combine clothing, complexion and height to recognize individuals by a humanoid. However, such information may not be easy to obtain for other kind of robots which may not see the whole body. There is a lack of results presented for HRI real-applications, likely due to the unrestricted acquisition setup [38]. Recent advances in face recognition have introduced the use deep learning for such task in robotics. In this sense, single robots such as the one presented in Jiang and Wang [15] or TERESA [38] integrate FaceNet [30] embeddings to perform face characterization.

More recently, robots guiding humans have taken a step further and began to visually identify the humans they seek to guide. In the works of [11] and [9] a complex model to identify social interaction between the robot and the human that is being assisted was proposed and tested. In those works, the role of identifying reliably the specific person in an environment that is constantly changing due to the movement of both human and robot is key. AveRobot [25], was created to provide a challenging benchmark for face recognition in HRI across multiple floors, where different features of relevant databases are compared. The recordings provide audio and video using eight sensors in different locations across three different floors, posing a complex re-identification and verification scenario [24]. This dataset is used to evaluate the approach proposed in this paper.

Also, recently several solutions to noise and low-quality environments have been proposed. In [5], a novel approach to tackle face detection in low resolution images can found. In this work a Gunnar Farneback optical flow is used first to understand moving parts of the image, while afterwards, Haar Cascades and Local Binary Patterns are used as to detect face on those moving parts. Another recent method that tries to advance face re-identification from low quality images can be found in Apicella et al. [4]. This paper proposes the use of super-resolution techniques to enhance the image of the face through a sequence of frames in order to improve performance and reliability. The results presented in his paper indicate that this is a viable option to tackle this task. Finally, a survey on techniques devoted to further the research on face recognition on low quality images can be found in Li et al. [21].

3. Visual proposal description

Biometric verification from visual cues is usually based on facial information. Thus, face detection is applied, using eye and mouth locations to crop each facial sample, composing a set of samples for each clip. The samples are ranked to select a subset to model an identity. Next subsections summarize the whole procedure.

3.1. Face detection

AveRobot presents unrestricted illumination conditions making face detection challenging. In addition, the dataset contains video clips recorded by eight different sensors, with some of them capturing interlaced video. For those particular videos, the clip frames are pre-processed to reduce interlacing artifacts that affect negatively face detection and cropping. Basically, odd lines are removed, resizing the resulting image to the original image dimension applying a pixel nearest interpolation.

Once pre-processing is done, if necessary, face detection is performed. After evaluating different standard face detectors [17,18], MTCNN [41] was chosen given the acceptable speed, stability and robust answer for this scenario.

3.2. Face cropping

MTCNN face detector provides a face container, and also five facial landmarks (eyes, nose and mouth), being eye and mouth locations used to guide the face cropping step. No further alignment is adopted in the experiments below. Being X the difference in the x axis between the eye locations, and Y the difference in the y axis between the eyes and mouth landmarks, the face container is forced to have a dimension of $3 \cdot X \times 3 \cdot Y$ pixels, locating the eyes at 1/3 and 2/3 in the x axis, and similarly eyes and mouth at 1/3 and 2/3 in the y axis.

3.3. Detection quality

Once that the collection of cropped faces is obtained for a given clip, each detection is ranked combining the MTCNN detector confidence, $conf_{det}$. and two Image Quality Assessment (IQA) criteria. The first one, *brisque*, is based on Blind/Referenceless Image Spatial Quality Evaluator (brisque) [27] that provides a value based on the image luminance. The second one, *cont*, introduces contour statistics assuming that typically a good contrasted facial image should contain a higher number of edge pixels. Therefore, after applying a Canny operator the number of edge pixels is divided by the total number of sample pixels. For two samples with identical $conf_{det}$, the one with larger IQA would be selected in first place. For the experimental setup used below, the final confidence value is calculated as follows:

$$confidence = X1 \cdot conf_{det} + X2 \cdot brisque + X3 \cdot cont \quad (1)$$

where $X1 + X2 + X3 = 1$ and $conf_{det}, brisque, cont \in [0, 100]$

3.4. Sample selection

After the previous steps, each face included in the collection of cropped faces of a given clip is characterized by a *confidence* value. At this point, a number of facial samples are chosen according to their *confidence*. The proposal assumes that samples with the same confidence are quite similar and likely located quite close in time. Therefore, the selection within the video clip is done forcing a difference in the confidence, to avoid the inclusion of rather similar captures and increase the intra-class variability. The objective is to extract the *max_detection* samples with highest quality, but not identical from the clip. In the experiments below, $max\_detection = 4$, but later just one of them is randomly taken to compute distances. Examples of selected and non-selected faces can be seen in Fig. 1. The samples selection is done according to the following steps:

1. The clip maximum *confidence* is obtained.
2. Samples are clustered according to their confidence value, using a step value of 0.00001, and sorted in descendent order from the maximum value.
3. The attention is given to the first cluster, i.e. with largest confidence.
4. While the number of collected samples is lower than *max_detection*:
   (a) A sample is taken randomly from the current cluster.
   (b) Next cluster in descendant order is taken.

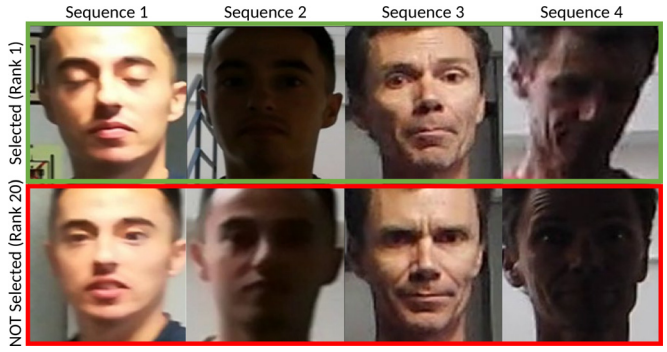

Fig. 1. Sample selection. **Top row:** Selected image from a specific sequence, the image shown is ranked as the best image by the face detector (highest confidence). **Bottom row:** Example of non selected image from a specific sequence, the image shown is ranked as the 20th image by the face detector (lowest confidence).

(c) When the last cluster is reached without reaching *max_detection* samples, clusters rejected samples are reconsidered, selecting those with highest confidence.

3.5. Descriptors

Each selected sample is used to feed a pre-trained neural network that computes face embeddings. This step encloses a face sample resize to match the net expected input dimensions. In some preliminary evaluations experimental setups, two of them have been explored: FaceNet which is based on an Inception ResNet v1 architecture [30], and the dataset VGGFace2 that is used to fine-tune a ResNet50 originally trained on MS-Celeb-1M [6]. Those preliminary experiments, not included here given the lack of space, reported a significantly better performance of the second approach, and therefore, VGGFace2 is used in the experiments presented below. Euclidean distance is later adopted in the visual-based verification matching.

4. Audio proposal description

The speaker verification, see Fig. 2, is tackled through three different stages. Raw audio is first pre-processed using a noise suppression technique. Secondly, a voice activity detection algorithm extracts the utterance speaker audio. Finally, the speaker audio is used in a feed-forward deep neural network for speaker verification purposes.

4.1. Noise suppression

The presence of highly non-stationary noise conditions can be considered problematic. In our sense, the detected noise fluctuations over short time scales can contaminate the audio signal. In our proposal, the RNNoise approach is used to address the noise suppression process. RNNoise is a technique to real-time full-band speech enhancement proposed by Valin [37]. This approach combines DSP-based techniques with deep learning. It consists of a conventional pitch filter and several hidden layers of a deep neural network. Traditional noise suppression algorithms make use of three different stages to tackle the problem: (i) voice activity detection, (ii) noise spectral estimation and iii) spectral subtraction. However, RNNoise uses a three-layer recurrent neural networks (RNN) instead of these three stages. RNN layers are organized in cascade, where each RNN layer is the input for the successive RNN layer. The input of the network is the frequency spectrum features of each frame, and the output is the frequency bands gain. Each original audio is converted to single-channel, 16-bit streams at a 48 kHz sampling rate in order to fit the RNNoise input. The band gain (g_b) is defined as follows:

$$g_b = \sqrt{\frac{E_s(b)}{E_x(b)}} \tag{2}$$

where $E_s(b)$ and $E_x(b)$ are the energy of the clean (ground truth) speech and the energy of the input (noisy) speech of the frequency band b respectively [37]. By observing Eq. (2), one may infer that the clean speech energy can be computed by adjusting a gain value (in the range [0, 1]) that multiplies the noisy speech energy.

4.2. Voice activity detection

Once the noise suppression has been handled, the next stage is to extract the speech regions of an utterance which are the most effective for speaker discrimination. If a large number of non-speech frames are considered for classification purposes, they can corrupt the decision process and hence significantly reduce the performance of our proposal. This second stage is known in literature as voice activity detection (VAD).

Although there is an extensive research on different VAD techniques, speaker verification and VAD techniques have been largely developed independently from each other [16]. As Jung stated, research on the use of VAD in the speaker verification context is very limited. Voice Onset Time (VOT) is defined as the period between the release of a plosive and the onset of vocal cord vibrations in the production of the following sound [32]. In other words, onset detection (or segmentation) is the means by which we can divide a signal into smaller units of sound. Several works in literature address the VAD problem trough the VOT. An interesting technique consists of backtracking detected onset events to the nearest preceding local minimum of an energy function. This technique basically rolls back the timing of detected onsets from a detected peak amplitude to the preceding minimum. This is very useful when using onsets to determine slice points for segmentation [14]. We make use of the onset backtracking technique to generate audio clusters from the input signal. Hence, given an audio signal $s[n]$, the energy function considered as input for the backtracking algorithm is defined as the root mean square (RMS) of the short-time Fourier transform:

$$E = \sqrt{\frac{1}{N}\sum_{i}^{N} |STFT\{s[i]\}|^2} \tag{3}$$

where *STFT* stands for the short-time Fourier transform. This transform divides a longer time signal into shorter segments of equal length and then computes the Fourier transform separately on each shorter segment. Once the backtracking technique is computed, the most suitable cluster is automatically selected depending on the density of peaks and the cluster location within the signal. This technique allow us to obtain a speech region of an utterance.

4.3. Audio classification

The last stage seeks to determine the identity of a speaker from the pre-processed audio. Usually, there are two common speaker recognition tasks; speaker identification [7,39] and speaker verification [20,33]. A speaker utterance is required to accomplish both tasks.

Traditionally, embedding methods have been used to map utterances into a low-dimensional feature space where distances correspond to speaker similarity. In this regard, i-vectors have been used to model inter-speaker variability [39,40]. A high-dimensional sample can be converted into a single low-dimensional i-vector that encodes speaker identity.

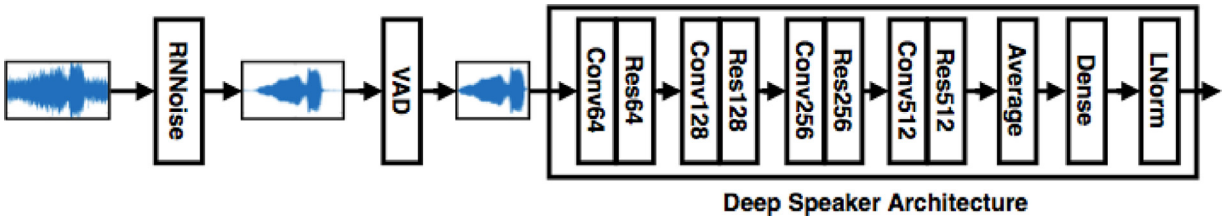

Fig. 2. The proposed audio pipeline. During the first stage, the audio signal is cleaned and the speaker utterance is extracted. Then, the frame level signal is processed through a pre-trained deep neural network [19]. Each of the three stacked ConvNet-ResNet blocks has an identical structure, and the skip connection is the identity mapping of x. The activation function considered is the clipped rectified linear (ReLU) function.

In this work, we used a pre-trained network known as Deep Speaker (see Fig. 2) proposed by Li et al. [19]. These authors used a ConvNet-ResNet deep architecture for frame level audio feature extraction. ResNet is composed of a number of stacked residual blocks. Each of these blocks contains direct links between the lower block outputs and the higher block inputs [12]. Afterwards, an average layer converts the frame-level input to an utterance-level speaker representation. Finally, the dense layer and a length normalization layer map the temporally-pooled features to a speaker embedding.

In terms of loss function, the cosine similarity has been proven as an efficient score function for speaker verification tasks [13,19]. Consequently, a triplet loss is considered as loss function. It operates on pairs of embeddings, by maximizing the cosine similarities of embedding pairs from the same speaker, and minimizing those from different speakers.

5. Multimodal classification: fusion of face and audio

Both biometric inputs described in Sections 3 and 4 provide a multimodal scenario. Hence, it is possible to combine the evidence presented by these biometric sources in order to verify the identity of an individual. It seems intuitive that by using more information results will improve and, as we will show in the experimental section, this is the case for the task at hand. It is for such a reason that we have performed several approaches that seek to take advantage of both biometric inputs. In this work, two different fusion strategies are described: one based in combing the biometric inputs at the feature vector level (Multimodal Feature Fusion), and another, based on combining the scores obtained after matching the feature vectors (Multimodal Score Fusion).

The first fusion strategy is obtained by performing a feature-level fusion. In this work, feature level fusion is accomplished by concatenating the feature sets obtained from the audio and images. Let $\mathbf{X_{audio}} = \{x_1, x_2, \ldots, x_m\}$ and $\mathbf{Y_{face}} = \{y_1, y_2, \ldots, y_n\}$ denote feature vectors ($\mathbf{X_{audio}} \in R^m$ and $\mathbf{Y_{face}} \in R^n$) representing information extracted via two different sources. The purpose is to combine both feature sets in order to yield a new feature vector $\mathbf{Z}$. The fused feature vector takes the following form $\mathbf{X_{fusion}} = \{x_1, \ldots, x_m, y_1, \ldots, y_n\}$. Then, as seen in Section 6, different approaches will perform a feature selection on the resultant feature vector.

The second fusion strategy adopted is a multimodal score-level fusion. In this case, the matching scores for each biometric modality are combined in order to provide a final verification decision. Hence, after both audio and face distances have been computed, a score-level weighted data fusion is performed according to the following rules:

$$score = score_{audio} \cdot w_{audio} + score_{face} \cdot w_{face} \tag{4}$$

where w_{audio} and w_{face} denote the audio and face weights respectively. Weights values are in the range $[0, 1]$ where $w_{audio} + w_{face} = 1$. In this regard, an experimental evaluation took place through a grid search considering different values for these weights. In all our experiments a weight distribution of $w_{audio} = 0.3$ and $w_{face} = 0.7$ has been empirically found to be the best.

6. Experimental evaluation

To evaluate the performance of both possible inputs to the robotic system we have used a dataset created for such purpose, the AveRobot dataset [25]. As briefly mentioned above, this dataset has showed to be more challenging than the previously existing as it presents many issues with lack of illumination in corridors. It also presents a wide array of identities, cameras and different locations within the same building. Such characteristics were selected to cover many possible situations in which a robot could find itself. The metric used for all experiments on the dataset is the commonly used Equal Error Rate (EER). This metric is specially suited to measure the performance of verification systems as it is the point where the false rejection rate (FRR) and the false acceptance rate (FAR) meet.

The experiments within the AveRobot dataset have been performed in the same way as in the original work [25] in order to create a fair comparison baseline. For each test dataset 40 identities are selected randomly. For each of those identities a list of 20 positive identity pairs (genuine pairs) and 20 false identity pairs (impostor pairs) are created. We then estimate the ability of each method to differentiate between genuine and impostor pairs using the EER metric as commented before. This experiment is performed 20 times using random identities in each iteration. All methods have been tested using the same random identities.

In the solutions we propose to all three combinations of voice and face biometric verification we apply dimensionality reduction to the embeddings. This serves two purposes, the first, to smooth the neural network response over the training samples to allow for a better generalization, and second, to allow for a better fusion of both face and voice inputs when performing Feature Fusion. Given the disparity in dimensionality, performing dimensionality reduction helps the combination of both by purging the less informative parts of each embedding. In our experiments we used Truncated Singular Value Decomposition (TSVD) [2] and Uniform Manifold Approximation and Projection (UMAP) [1]; apart from Principal Component Analysis (PCA), which was our final selection based on the obtained results. We have observed no benefit from using other methods and performance is fairly similar.

6.1. Face verification results

When performing face verification experiments all the dimensionality reduction techniques reported similar results: UMAP (14% - 20 components), PCA (13.38% - 100 components) and TSVD (13.48% - 100 components). These results are slightly better than considering just the raw embeddings (14.76% - 2048 components). However, it can be seen that compared to previous results from [24] a great improvement is obtained by pruning the chosen frames (see Table 1). Having stable and varied inputs helps

Table 1
Verification results on Averobot-EER. In this table we compare our approach with the work developed in Marras et al. [24].

Modal approach	Loss - Author	EER
Face Verification	Margin - Marras	45.43%
	Center - Marras	45.47%
	Ring - Marras	42.63%
	Softmax - Marras	38.20%
	ResNet50 - Ours	14.76%
	ResNet50 + PCA - Ours	**13.38%**
Voice Verification	Margin - Marras	44.58%
	Center - Marras	42.33%
	Ring - Marras	43.27%
	Softmax - Marras	41.58%
	Softmax - Ours	38.54%
	Triplet - Ours	34.05%
	Triplet + PCA - Ours	**32.07%**
Multimodal Verification	Margin - Marras	37.08%
	Center - Marras	32.80%
	Ring - Marras	34.40%
	Softmax - Marras	33.05%
	Feature Fusion - Ours	13.35%
	Score Fusion - Ours	**12.22%**

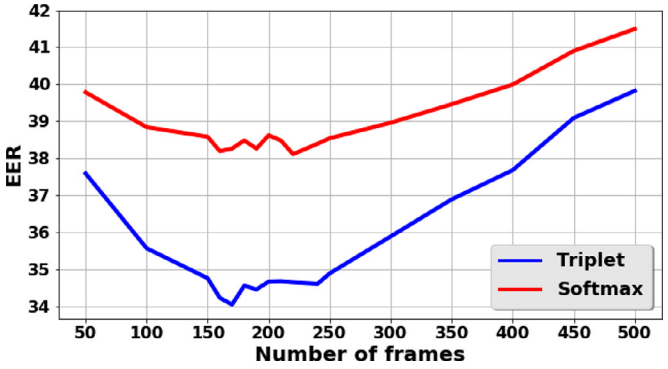

Fig. 3. Uni-modal voice verification results on Averobot-EER. A hundred frames are approximately one second.

the robot make better decisions when determining the identity of the person who is in front of it. By taking into account the context of the robot such an improvement is possible. In this particular case, the smoothing of the embedding space by performing a PCA dimensionality reduction lowers the error rate from 14.76% to 13.38%.

6.2. Audio verification results

As can bee seen in Fig. 3, the number of considered frames plays a key role. Accordingly to Section 4.2, the VAD ensures the location of speech frames at the beginning of each file. Otherwise, if not enough speech frames are detected, then the neural network may not be able to detect the speaker. It is justified to truncate all the audio samples somehow because we want all of them to have the same number of frames. For instance, let's consider 200 frames per sample. On the one hand, for the audio samples shorter than 200 frames, we pad them with silence. On the other hand, for the audio samples larger than 200 frames, we truncate them. Fig. 3 shows this trade-off between the number of frames, finding the best results when 160 frames are considered. The considered techniques reported analogous results: UMAP (33.02% - 20 components), PCA (32.07% - 200 components) and TSVD (33.96% - 80 components). Again these results are quite similar between them but better than considering just the raw embeddings (34.05% - 512 components).

In Table 1 the EER results can be seen for our proposed audio identity verification approach. In this case, it can also be seen that

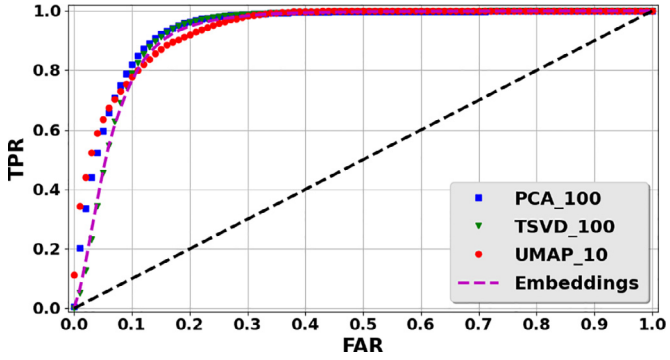

Fig. 4. ROC curves for score fusion related to the best performing approach for each technique.

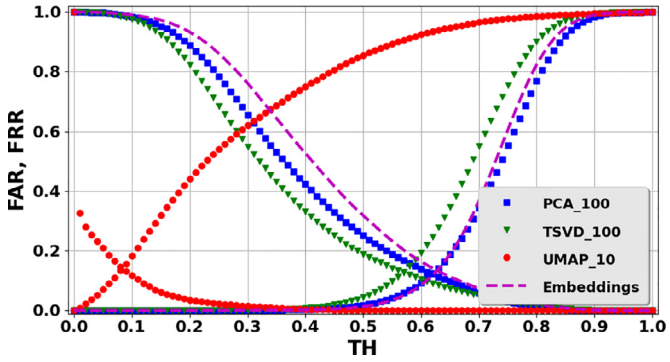

Fig. 5. FAR/FRR curves for score fusion related to the best performing approach for each technique.

paying attention to only the parts of the speech that are of interest nearly a 10% improvement is obtained between the best of the compared baselines and our best approach (PCA triplet). We show results of both our chosen neural network trained with a softmax and a triplet loss. In the reported results, the triplet loss performs much better.

6.3. Fusion verification results

The third experiment shows the results achieved in this work by combining both biometric inputs to achieve better global results. As we stated in Section 5, two different fusion strategies are considered in our work. The feature-level fusion scheme adopted is based on the concatenation of the two embedding vectors, while the score-level fusion scheme exploits variable weights for the voice and face components.

On the one hand, the feature fusion experiment reported the following rates: raw embeddings (15.02% - 2560 components), UMAP (14.36% - 10 components), PCA (13.35% - 160 components) and TSVD (14.32% - 190 components). On the other hand, the score fusion experiment reported the following rates: raw embeddings (13.33% - 2560 components), UMAP (13.39% - 10 components), PCA (12.2% - 160 components) and TSVD (12.58% - 160 components). In all our score fusion experiments a weight distribution of $w_{audio} = 0.3$ for audio biometric signal and $w_{face} = 0.7$ for face biometric signal has been empirically found to be the best. Table 1 shows the results for each fusion strategy. Both fusion strategies outperform all uni-modal approaches, being the best fusion strategy the score-level approach. We also report the full ROC curves (Fig. 4) for all embedding approaches used in our tests. Moreover, Fig. 5 shows FAR/FRR curves for the best embedding approaches when the score fusion is considered. It can be appreciated that perfor-

mance is quite similar among the considered approaches. For all dimensionality reduction methods used several number of components were tested. For all methods components in a range from 10 to 200 were tested, the number of components reported in this paper is the one with the highest experimental results for each.

7. Conclusions

We have showed in this work that by understanding the context and the task in which identity verification is performed remarkable improvements in performance can be achieved. By acknowledging that a robot can take a second, and even a third look, to the person the overall performance of identity verification from either audio or faces can be greatly boosted. By performing a careful selection of the regions of interest for voice and face identity verification complex real life scenarios, like the ones handled in this paper, can be solved. In this worked we have also demonstrated that by combining both face and voice biometric inputs a more robust approach can be achieved. It is the use of multiple inputs what can make a robot capable of handling the complex scenarios that it will encounter on a daily basis. We believe that with solutions that report high levels of robustness, like the ones presented in this work, robots can begin to helps us in an increasing manner.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is partially funded by the ULPGC under project ULPGC2018-08, by ACIISI and ULPGC under project CEI-2018-4, the Spanish Ministry of Economy and Competitiveness (MINECO) under project RTI2018-093337-B-I00, and by FEDER funds under project ProID2020010024. Adrian Penate-Sanchez is funded by a Beatriz Galindo grant from the Ministy of Education of Spain.

References

[1] L. McInnes, J. Healy, J. Melville, UMAP: uniform manifold approximation and projection for dimension reduction, 2018, arXiv:1802.03426.
[2] N. Halko, P.-G. Martinsson, J. A. Tropp, Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions, 2009, arXiv:0909.4061.
[3] M.K. Al-Qaderi, A.B. Rad, A multi-modal person recognition system for social robots, Appl. Sci. 8 (3) (2018) 387.
[4] A. Apicella, F. Isgró, D. Riccio, Improving face recognition in low-quality video sequences: single frame vs. multi-frame super-resolution, in: International Conference on Image Analysis and Processing., 2017.
[5] R.J. Cardenas T., C.A.B. Castañón, J.C.G. Cáceres, Face detection on real low resolution surveillance videos, in: Proceedings of the 2nd International Conference on Compute and Data Analysis, 2018, pp. 52–59.
[6] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, Vggface2: adataset for recognising faces across pose and age, 2017, arXiv:1710.08092.
[7] R. Chakroun, M. Frikha, L. Beltaïfa zouari, New approach for short utterance speaker identification, IET Signal Proc. 12 (2018) 873–880.
[8] F. Faber, M. Bennewitz, C. Eppner, A. Görög, C. Gonsionr, D. Joho, M. Schreiber, S. Behnke, The humanoid museum tour guide Robotinho, in: IEEE International Symposium onRobot and Human Interactive Communication (RO-MAN), 2009, pp. 891–896.
[9] G. Ferrer, A.G. Zulueta, F.H. Cotarelo, A. Sanfeliu, Robot social-aware navigation framework to accompany people walking side-by-side, Auton. Robots 41 (2017) 775–793.
[10] D.F. Glas, S. Satake, F. Ferreri, T. Kanda, H. Ishiguro, N. Hagita, The network robot system: enabling social human-robot interaction in public spaces, Int. J. Human-Robot Interact. 1 (2) (2012) 5–32.
[11] A. Goldhoorn, A. Garrell, R. Alquézar, A. Sanfeliu, Searching and tracking people with cooperative mobile robots, Auton Robots 42 (2018) 739–759.
[12] Z. He, X. Li, Z. Zhang, Y. Zhang, J. Xiao, X. Zhou, Structure-aware slow feature analysis for age estimation, IEEE Signal Process. Lett. 23 (2016) 1702–1706.
[13] G. Heigold, I. Moreno, S. Bengio, N. Shazeer, End-to-end text-dependent speaker verification, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5115–5119.
[14] T. Jehan, Creating Music by Listening, 2005 Ph.D. thesis. Massachusetts Institute of Technology, School of Architecture and Planning.
[15] W. Jiang, W. Wang, Face detection and recognition for home service robots with end-to-end deep neural networks, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, 2017, pp. 2232–2236.
[16] Y. Jung, Y. Choi, H. Kim, Self-adaptive soft voice activity detection using deep neural networks for robust speaker verification, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 365–372.
[17] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1867–1874.
[18] D.E. King, Max-margin object detection, 2015, arXiv:1502.00046.
[19] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, Z. Zhu, Deep speaker: an end-to-end neural speaker embedding system, CoRR abs/1705.02304(2017).
[20] J. Li, M. Sun, X. Zhang, Y. Wang, Joint decision of anti-spoofing and automatic speaker verification by multi-task learning with contrastive loss, IEEE Access 8 (2020) 7907–7915.
[21] P. Li, L. Prieto, D. Mery, P. Flynn, Face recognition in low quality images: a survey, CoRR abs/1805.11519(2018).
[22] J. López, D. Pérez, M. Santos, M. Cacho, Guidebot. A tour guide system based on mobile robots, Int. J. Adv. Rob. Syst. 10 (11) (2013) 1–14.
[23] J. López, D. Pérez, E. Zalama, J. Gomez-Garcia-Bermejo, Bellbot - a hotel assistant system using mobile robots, Int. J. Adv. Robot. Syst. 10 (1) (2013) 1–11.
[24] M. Marras, P.A. Marín-Reyes, J. Lorenzo-Navarro, M. Castrillón-Santana, G. Fenu, Deep multi-biometric fusion for audio-visual user re-identification and verification, Pattern Recognition Applications and Methods. ICPRAM 2019, 2020.
[25] M. Marras, P. Marín-Reyes, J. Lorenzo-Navarro., M. Castrillón-Santana., G. Fenu., AveRobot: an audio-visual dataset for people re-identification and verification in human-robot interaction, ICPRAM, 2019.
[26] E. Martinson, W. Lawson, G. Trafton, Identifying people with soft-biometrics at fleet week, in: Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction, 2013.
[27] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.
[28] O. Parra, I. Rodriguez, E. Jauregi, E. Lazkano, T. Ruiz, Gidabot: a system of heterogeneous robots collaborating as guides in multi-floor environments, Int. J. Serv. Robot. 12 (2019) 319–332.
[29] I. Rodriguez, U. Zabala, P. Marín-Reyes, E. Jauregi, J. Lorenzo-Navarro, E. Lazkano, M. Castrillón-Santana, Personal guides: heterogeneous robots sharing personal tours in multi-floor environments, Sensors 20 (9) (2020).
[30] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.
[31] M. Shiomi, T. Kanda, H. Ishiguro, N. Hagita, Interactive humanoid robots for a science museum, IEEE Intell. Syst. 22 (2) (2007) 25–32.
[32] R. Singh, J. Keshet, D. Gencaga, B. Raj, The relationship of voice onset time and voice offset time to physical age, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, 2016, pp. 5390–5394.
[33] K. Sriskandaraja, V. Sethu, E. Ambikairajah, H. Li, Front-end for antispoofing countermeasures in speaker verification: scattering spectral decomposition, IEEE J. Sel. Top. Signal Process. 11 (2017) 632–643.
[34] L. Susperregi, I. Fernandez, A. Fernandez, S. Fernandez, I. Maurtua, I.L. de Vallejo, Interacting with a robot: a guide robot understanding natural language instructions, in: Ubiquitous Computing and Ambient Intelligence, Springer, 2012, pp. 185–192.
[35] S. Thrun, M. Bennewitz, W. Burgard, A.B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Schulte, D. Schulz, Minerva: asecond-generation museum tour-guide robot, International Conference on Robotics and Automation, IEEE, 1999, pp. 1999–2005.
[36] D. Troniak, J. Sattar, A. Gupta, J.J. Little, W. Chan, E. Calisgan, E. Croft, M. Van der Loos, Charlie rides the elevator–integrating vision, navigation and manipulation towards multi-floor robot locomotion, Computer and Robot Vision (CRV), 2013 International Conference on, IEEE, 2013, pp. 1–8.
[37] J. Valin, A hybrid DSP/deep learning approach to real-time full-band speech enhancement, in: 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), 2018, pp. 1–5.
[38] Y. Wang, J. Shen, S. Petridis, M. Pantic, A real-time and unsupervised face re-identification system for human-robot interaction, Pattern Recognit. Lett. 128 (2019) 559–568.
[39] L. Xu, Z. Yang, L. Sun, Simplification of i-vector extraction for speaker identification, Chin. J. Electron. 25 (2016) 1121–1126.
[40] S. Yao, R. Zhou, P. Zhang, Y. Yan, Discriminatively learned network for i-vector based speaker recognition, Electron. Lett. 54 (2018) 1302–1304.
[41] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (2016) 1499–1503.