AveRobot: An Audio-visual Dataset for People Re-identification and Verification in Human-Robot Interaction

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*

[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
{mirko.marras, fenu}@unica.it, {javier.lorenzo, modesto.castrillon}@ulpgc.es, pedro.marin102@alu.ulpgc.es

Keywords: Face-voice Dataset, Deep Learning, People Verification, People Re-Identification, Human-Robot Interaction.

Abstract: Intelligent technologies have pervaded our daily life, making it easier for people to complete their activities. One emerging application is involving the use of robots for assisting people in various tasks (e.g., visiting a museum). In this context, it is crucial to enable robots to correctly identify people. Existing robots often use facial information to establish the identity of a person of interest. But, the face alone may not offer enough relevant information due to variations in pose, illumination, resolution and recording distance. Other biometric modalities like the voice can improve the recognition performance in these conditions. However, the existing datasets in robotic scenarios usually do not include the audio cue and tend to suffer from one or more limitations: most of them are acquired under controlled conditions, limited in number of identities or samples per user, collected by the same recording device, and/or not freely available. In this paper, we propose *AveRobot*, an audio-visual dataset of 111 participants vocalizing short sentences under robot assistance scenarios. The collection took place into a three-floor building through eight different cameras with built-in microphones. The performance for face and voice re-identification and verification was evaluated on this dataset with deep learning baselines, and compared against audio-visual datasets from diverse scenarios. The results showed that *AveRobot* is a challenging dataset for people re-identification and verification.

1 INTRODUCTION

Humans have been expecting the integration of intelligent robots in their daily routine for many years. In this vision, one of the main applications receiving great attention involves the use of assistance robots. The literature describes a wide range of robots acting individually as tour guides since the late 90's (Thrun et al. 1999; Domínguez-Brito et al. 2001). More recently, the focus was moved to the social interaction between robots and humans. Integrating natural language processing and semantic understanding had a great success in different areas to this end (e.g., Boratto et al. 2017; Shiomi et al. 2007). In the robotics context, joined with the contribution of path optimization theory (Mac et al. 2017; Fenu and Nitti 2011), they made possible to improve the cognitive and mobility capabilities of robots while guiding or assisting visitors (Susperregi et al. 2012). Furthermore, some robots were equipped with a wheeled platform to reduce mobility constraints (Faber et al. 2009), while others showed pro-active capabilities with the visitors for completing assigned tasks (Rosenthal et al. 2010). In the same direction, under populated environments, multiple interactive robots acting as guides cooperated by sharing users' profiles and tour information (Trahanias et al. 2010; Hristoskova et al. 2012).

In the latter dynamic scenario, multiple and likely different robots act as coordinated assistants for any visitor and need to cooperate with each other. These configuration can avoid challenging and dangerous situations of multi-floor movements (e.g, Troniak et al. 2013; López et al. 2013a,b) and reduce the complexity required for implementing navigation. However, for certain tasks, this setup imposes the interchanging of descriptors about the visitors among a group of heterogeneous robots. For instance, this is required for guiding a person when s/he moves from one floor to another, so that the receiving robot can pro-actively identify the assisted person among other visitors. Recognizing the redirected person is an expected capability for the receiving robot. This can be viewed both as a re-identification or verification task.

Embedded applications often use the face features to establish the identity of a person of interest (Cruz et al. 2008; Barra et al. 2013; Taigman et al. 2014). Unfortunately, the face may not offer enough information in many scenarios due to variations in pose, illumination, resolution and recording distance. Other biometric modalities, such as the voice, may improve the recognition performance (Ouellet et al. 2014; Nagrani et al. 2017), but they also suffer from environmental noise or distance to the microphone. Considering two or more biometric modalities generally tends to make the system more reliable due to the presence of multiple independent pieces of evidence (Fenu et al. 2018; Fenu and Marras 2018; Barra et al. 2017). However, the existing datasets collected in Human-Robot Interaction (HRI) scenarios usually include no audio cue and tend to suffer from one or more limitations: they are obtained under controlled conditions, composed by a small number of users or samples per user, collected from the same device, and not freely available.

The contribution of this paper is twofold. The first one includes a pipeline for creating an audio-visual dataset tailored for testing biometric re-identification and verification capabilities of robots under a multi-floor cooperation scenario. By using tripods equipped with multiple recording devices and semi/fully-automated processing scripts, we simulate different robot acquisition systems and reduce the human intervention during the dataset construction. We leverage this pipeline to collect *AveRobot*, a multi-biometric dataset of 111 participants vocalizing short sentences under robot assistance scenarios. The collection took place into a three-floor building by means of eight recording devices, targeting various challenging conditions. The second contribution involves the investigation of different techniques for training deep neural networks on face and spectrogram images extracted directly from the frames and the raw audios, and the comparison of the performance on this new dataset against the performance the same techniques obtain on other traditional audio-visual datasets recorded on different scenarios. We provide baselines for face and voice re-identification and verification tasks to assess the relevance and the usefulness of our dataset. The results show that the dataset we are providing appears as challenging due to the uncontrolled conditions. The *AveRobot* dataset is publicly available at **http://mozart.dis.ulpgc.es/averobot/**.

The paper is organized as follows. Section 2 depicts the related work, while Section 3 describes the proposed pipeline and the resulting dataset. Section 4 shows the results and Section 5 concludes the paper.

2 RELATED WORK

In this section, we discuss various literature contributions relevant to the creation of the dataset presented in this paper. First, we describe traditional and deep learning methods for biometric recognition; then, we compare datasets used by previous works.

Traditional HRI Methods. The field of biometric recognition in HRI was dominated by techniques integrating hand-crafted features related to both hard biometrics, such as face, and soft biometrics, such as gender, age, and height (Cielniak and Duckett 2003; Cruz et al. 2008). The combination of audio-visual biometric features was leveraged by Martinson and Lawson (2011). Their method performed face recognition through basic neural networks and speaker recognition with Gaussian Mixture Models (GMMs). To increase the robustness under uncontrolled scenarios, Ouellet et al. (2014) combined face and voice identification with human metrology features (e.g., anthropometric measurements). Correa et al. (2012) modelled faces in the thermal and visual spectra for the same goal. In case of bad illumination, their approach relies on thermal information, while thermal and visual information complement each other in good illumination scenarios. Skeleton data was leveraged by Sinha et al. (2013) to detect gait cycles and compute features based on them. Feature selection and classification were performed with adaptive neural networks. Illumination-independent features (i.e., height and gait) were also used by Koide and Miura (2016). To manage the limitations of RGB and skeleton features in dealing with occlusions and orientation, Cosar et al. (2017) and Liu et al. (2017) presented RGB-D-based approaches using features from the body volume. The work proposed by Irfan et al. (2018) used a multi-modal Bayesian network for integrating soft biometrics together with the primary information provided by faces.

Deep Learning Methods. The recent widespread of deep learning in different areas (e.g., Boratto et al. 2016, Nagrani et al. 2017) has motivated the use of the neural networks as feature extractors combined with classifiers, as proposed in Wang et al. 2018b. For face recognition, backbone architectures rapidly evolved from AlexNet (Krizhevsky et al. 2012) to SENet (Hu et al. 2017). Deepface (Taigman et al. 2014) and its variations use a cross-entropy-based Softmax loss as a metric learning while training the network. However, Softmax loss is not sufficient by itself to learn features with large margin, and other loss functions were explored to enhance the generalization ability. For instance, euclidean distance based losses embed images into an euclidean space and reduce intra-

Table 1: The comparison of the existing datasets for biometric identification and verification in Human-Robot Interaction.

Dataset	Users	Is Public?	Devices/User	Videos/User	Visual Specs	Audio Specs
Correa et al. 2012	16-171	Yes	2	1-4	RGB + RGB-D	-
Munaro et al. 2014	50	Yes	1	5	RGB + RGB-D	-
Ouellet et al. 2014	22	No	1	3	RGB + RGB-D	16bit 48kHz
Liu et al. 2017	90	Yes	1	2	RGB + RGB-D	-
Wang et al. 2018b	26	No	1	1	Grayscale	-
Irfan et al. 2018	14	No	1	4	RGB	-
AveRobot (Ours)	**111**	**Yes**	**8**	**24**	**RGB**	**16bit 16kHz**

variance while enlarging inter-variance. Contrastive loss (Sun et al. 2015) and Triplet loss (Schroff et al. 2015) are commonly used, but sometimes they exhibit training instability. Center loss (Wen et al. 2016) and Ring loss (Zheng et al. 2018) are good alternatives. Angular/cosine-margin based losses were proposed to learn features separable through angular/cosine distance (Wang et al. 2018a). For speaker recognition, GMMs and i-vectors models were originally used on top of a low dimensional representation called Mel Frequency Cepstrum Coefficients (MFCCs) (Hansen and Hasan 2015). However, their performance degrades rapidly in real world applications. They focus only on the overall spectral envelope of short frames. This led to a shift from hand-crafted features to neural approaches trained on high dimensional inputs (Lukic et al. 2016;Nagrani et al. 2017).

HRI Datasets. Table 1 reports a representative set of datasets for people recognition in HRI scenarios. Correa et al. (2012) generated four different image databases to compare the use of visual and thermal information. To inspect RGB-D images, Munaro et al. (2014) collected a dataset of 50 different subjects, performing a certain routine of motions in front of a Kinect. The dataset includes synchronized RGB and depth images, persons segmentation maps and skeletal data taken in two different locations. In the same direction, Liu et al. (2017) collected a dataset including 90 users, each recorded during sitting and walking in two different rooms. Moreover, to enable audio-visual people recognition, Ouellet et al. (2014) used a Kinect camera to capture RGB images and a microphone array with eight channels to collect audios. The data were acquired for 22 participants sitting and moving their faces in different poses. Furthermore, Wang et al. (2018b) created a database recording the daily interactions between a robot and its users. More recently, Irfan et al. (2018) collected data from 14 participants over four weeks in the context of real-world applications for cardiac rehabilitation therapy with a personalized robot. Summarizing, existing datasets usually suffer from one or more of the following limitations: they are obtained under controlled conditions, limited in size or samples per user, collected by the same device, do not include audio, or not freely available. By contrast, our dataset has been designed to try to reduce those limitations.

3 THE PROPOSED DATASET

In this section, we introduce the proposed dataset, including the scenario, the statistics, the environmental setup and the collection pipeline adopted.

3.1 Collection Scenario

The data gathering conducted in this work involved acquiring audio-visual data of participants reproducing short sentences in front of recording devices in indoor environments. The resulting dataset is referred to as *AveRobot*. The main goal was to mimic a robot assistance scenario in a semi-constrained indoor environment, as often encountered in public buildings like universities or museums. More precisely, the data collection took place inside a three-floor office building. Considering that the problem was related to the robot sensory part, no real robots were necessary, but they were simulated through the use of various cameras and microphones similar to the ones integrated into robots. As a result, the interactions in each floor were recorded with different devices, simulating a total number of eight robot acquisition systems: two in the first floor, three in the second one, and three in the third one. Furthermore, as a person has two options to reach another floor (i.e. using the elevator or the stairs), the recordings were made at three locations for each floor: near the stairs, along the corridor, and outside the lift. Figure 1 provides sample faces detected in *AveRobot* videos. As the reader may expect, the use of different acquisition devices poses changes in image illumination, geometry and resolution, and sound quality. In fact, the acquisition was degraded with real-world noise, consisting of background chatter, laughter, overlapping speech, room acoustics, and there was a range in the quality of recording equipment and channel noise.

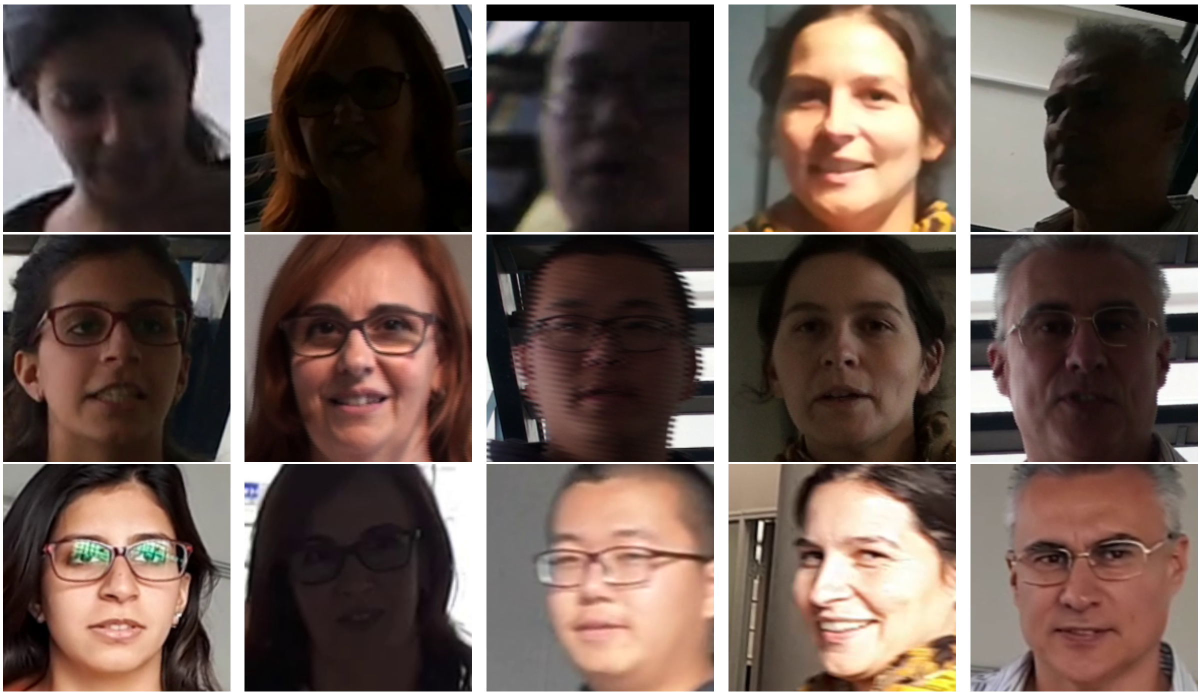


Figure 1: Samples from the dataset proposed in this paper. Each column corresponds to a specific participant and shows one acquisition per floor. The samples depict variations in pose, illumination, resolution and recording distance.

3.2 Collection Pipeline

To collect the proposed audio-visual dataset, we followed a pipeline whose steps are described as follows.

Step 1: Device Selection. Eight recording devices were selected to make up the dataset, each simulating a different robot acquisition system. Table 2 details their characteristics. It should be noted that the devices expose different peculiarities and they are similar to the sensors embedded in robots. Camera 1 and 7 tended to generate more blurred recordings. On the other hand, Camera 3 and 6 recorded videos using interlaced scan, differently from the progressive scan performed by the others.

Step 2: Environmental Setup. We grouped the devices per floor by considering their different type and various operational heights. Floor 0 hosted Camera 1 and 2 at a fixed height of 130*cm*, Floor 1 included Camera 3, 4 and 5 at a fixed height of 120*cm*, and Camera 6, 7 and 8 worked on Floor 3 at a fixed height of 150*cm*. Therefore, each floor hosted a smartphone camera, a compact camera and a video camera, except Floor 0. To assure that the recordings were done in similar conditions, tripods were used for compact and video cameras, while smartphone cameras were held by a human operator at the same height of the other devices. In most cases, we selected a recording height lower than a human because robots are typically not very tall (e.g., Pepper is 120*cm* height). The devices were configured with the highest possible resolution at a constant frame rate (25 *fps* for Camera 6 and 30 *fps* for the remaining cameras).

Step 3: User Recording. The identical recording procedure was repeated for each user of our dataset. Firstly, for each location, the user selected and memorized the sentence to be articulated, taken from a list of pre-defined sentences. Meanwhile, the devices were arranged in a position near the target location (i.e. stairs, corridor and lift). Then, the human operators switched on the corresponding devices at the same time, while the user approached the camera and reproduced the sentence in front of the capturing devices. In this way, at each location, the same speech was simultaneously recorded with two/three devices. The same process was repeated on each floor and location by selecting a different sentence. The overall process took between 6 and 10 minutes per user.

Step 4: Data Protection. After finishing the session, the user read and signed an appropriate agreement in order to respect the European data protection regulation. The information provided by the participant included but it was not limited to: her/his full name, the identification number, whether s/he authorizes or not to show their data as samples on research articles, and the signature. Gender, height and age were registered.

Step 5: Video Labelling. The videos were manually labelled to keep track of the participant identity, floor and location, the pronounced sentence and the recor-

Table 2: The specifications of the recording devices used for the dataset construction.

ID	Model	Type	Resolution	Fps	Format	Height (*cm*)	Floor
1	Casio Exilim EXFH20	Compact Camera	1280×720	30	AVI	130	0
2	Huawei P10 Lite	Smartphone Camera	1920×1080	30	MP4		
3	Sony HDR-XR520VE	Video Camera	1920×1080	30	MTS	120	1
4	Samsung NX1000	Compact Camera	1920×1080	30	MP4		
5	iPhone 6S	Smartphone Camera	1920×1080	30	MOV		
6	Sony DCR-SR90	Video Camera	720×576	25	MPG	150	2
7	Olympus VR310	Compact Camera	1280×720	30	AVI		
8	Samsung Galaxy A5	Smartphone Camera	1280×720	30	MP4		

ding device. To this end, each video was properly renamed by using the following convention: UserId-FloorId-LocationId-SentenceId-DeviceId. Moreover, a metadata file was created to save the personal information regarding each user: the assigned id, name, surname, gender, height, native language, age and approval status (i.e. if they authorized to publish their data in research articles). The anonymized version is made publicly available together with the dataset.

Step 6: Visual Post-Processing. First, all the videos were converted from the original video format to the target MP4 format. Then, the faces were detected and aligned with MTCNN (Zhang et al. 2016), resized to 224×224 pixels and stored as PNG images into a specific folder. Those frames with detected faces were extracted and saved, with original resolution, as PNG images into another folder. Each image was manually checked to remove false positives.

Step 7: Audio Post-Processing. Once that the audio was extracted from each video and stored as a WAV file, the silence part at the beginning and ending of the audio was removed through a semi-automated process which involved the Auditok[1] segmentation tool. Therefore, the resulting audios included only the part when the participant talks. Third, each audio was converted to single-channel, 16-bit streams at a 16kHz sampling rate. The related spectrograms were generated in a sliding window fashion using a Hamming window of width $25ms$ and step $10ms$ for each second of speech. As a result, multiple spectrograms were generated for each audio.

3.3 Dataset Statistics

The proposed dataset contains $2,664$ videos from 111 participants (65% male and 35% female) who vocalize different short sentences. The sentences were selected by the participant from a pre-defined set of 34 sentences tailored for a robot assistance scenario. The collected people span different ethnicities (e.g., Chinese and Indian), ages (avg. 27; std. 11; min. 18; max. 60), and heights (avg. $1.74m$; std. $0.10m$; min. $1.50m$; max. $1.92m$). Figure 2 depicts relevant distributions along the dataset. The gender, height, and age for each participant are also provided together with the videos. Each person was recorded in 3 locations (i.e. stairs, floor and lift) for each one of the 3 floors of the building. As mentioned above, 8 diverse recording devices were leveraged during the collection to simulate the robot acquisition systems. The recording devices assigned to the same floor worked simultaneously. Thus, the dataset comprises 24 videos per user:

- 1st Floor: 2 (devices) $\times$ 3 (locations) = 6 videos.
- 2nd Floor: 3 (devices) $\times$ 3 (locations) = 9 videos.
- 3rd Floor: 3 (devices) $\times$ 3 (locations) = 9 videos.

The total length of the video resources provided by the proposed dataset is $5h$ $17min$, occupying $21.8GB$. Each participant is represented by more than $3min$ of videos, each lasting around $7s$. It should be noted that each video includes three phases: (i) when the person is approaching to the devices, (ii) when s/he speaks in front of them, and (iii) when s/he leaves the scene. Hence, looking only at the face content, each video contains around 127 frames with a detected face and each user is represented by over $3,000$ detected faces. The total number of detected faces is $338,578$, occupying $18.0GB$. On the other hand, looking at the voice content, each video contains around $3s$ of speech and each user is represented by over $1m$ of content. The total length of the voice data is around $1h$ $40min$, occupying $283MB$.

4 EXPERIMENTS

In this section, we evaluate the wildness of *AveRobot* by conducting a number of baseline experiments. First, we detail the implementation of the neural network architecture and the resulting loss functions selected as baselines. Then, the experimental protocols are described for both people re-identification and verification. Figure 3 provides an overview of our experimental and evaluation methodology. Finally, we

[1] https://github.com/amsehili/auditok

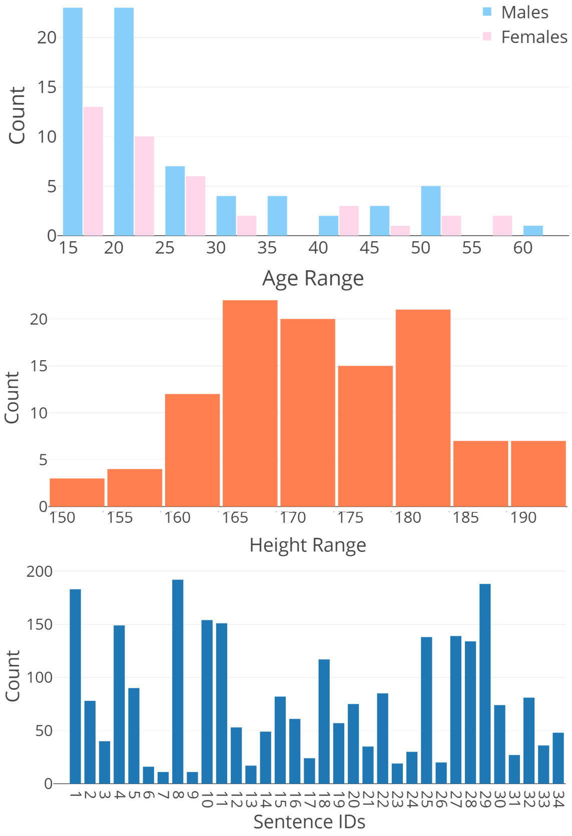


Figure 2: The dataset statistics about the gender per age distribution (**top**), the user height distribution (**center**), and the pronounced sentence distribution among videos (**bottom**).

compare the performance on *AveRobot* and other traditional audio-visual datasets.

4.1 Training and Testing Datasets

To the best of our knowledge, no public and large audio-visual dataset has been proposed for face and voice re-identification and verification in HRI scenarios. As a result, we leveraged traditional audio-visual datasets for training the baselines and we tested them not only on *AveRobot*, but also on datasets from diverse audio-visual contexts. First, this choice enables the computation of state-of-the-art deep learning baseline scores on *AveRobot*. Second, it can be possible to observe how the baselines differently perform on *AveRobot* and other traditional audio-visual datasets, giving an overview of the challenging level of *AveRobot*. The audio-visual datasets we included were divided in one training dataset and several testing datasets to replicate a cross-dataset setup.

Training Dataset. VoxCeleb Train Split is an audio-visual speaker identification and verification dataset collected by Nagrani et al. (2017) from Youtube, including 21,819 videos from 1,211 identities. It is the most suited for training a deep neural network due to the wide range of users and samples per user.

Testing Dataset 1. VoxCeleb Test Split is an audio-visual speaker identification and verification dataset collected by Nagrani et al. (2017) from Youtube, embracing 677 videos from 40 identities.

Testing Dataset 2. MOBIO is a face and speaker recognition dataset collected by McCool et al. (2012) 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 Chowdhury et al. (2018) under semi-controlled indoor surveillance scenarios, including 2,260 videos from 50 identities.

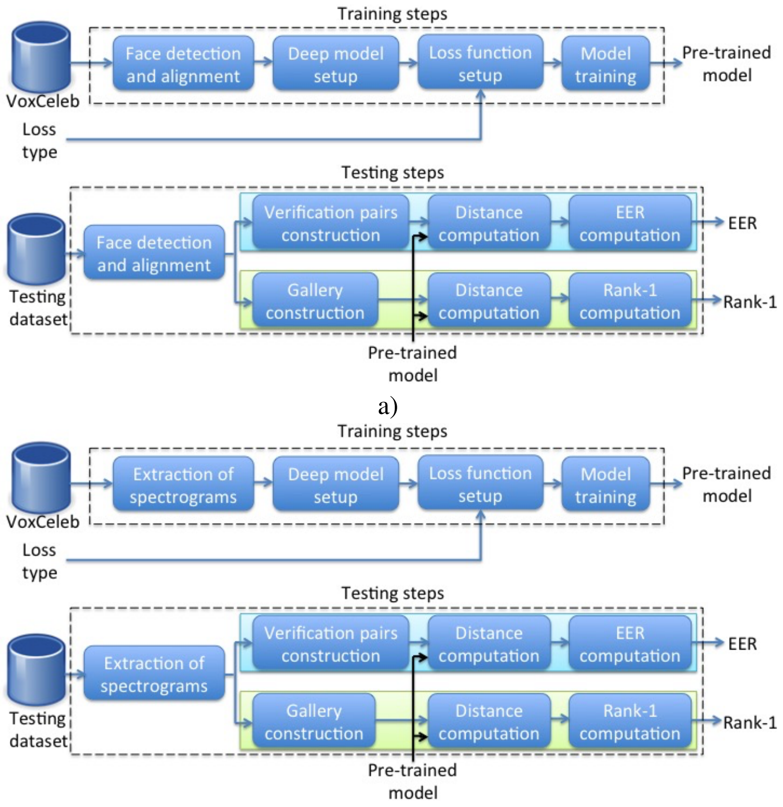
Testing Dataset 4. The dataset proposed in this paper, *AveRobot*, is an audio-visual biometric recognition dataset collected under robot assistance scenarios, including 2,664 videos from 111 identities.

4.2 Evaluation Setup

Face Input Features. As mentioned above, each frame is analyzed in order to detect the face area and landmarks through MTCNN (Zhang et al. 2016). The five facial points (two eyes, nose and two mouth corners) are adopted to perform the face alignment. The faces are then resized to 112 × 112 pixels in order to fit in our model 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 deep neural network. It should be noted that the face image size considered at this step differs from the one used during the visual post-processing of our dataset due to efficiency reasons. Thus, it was applied to all the considered datasets, so that the same face image size was maintained for all of them.

Voice Input Features. Each audio is converted to single-channel, 16-bit streams at a 16kHz sampling rate for consistency. The spectrograms are then generated in a sliding window fashion using a Hamming window of width 25ms and step 10ms. This gives spectrograms of size 112 × 112 for one second 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 neural network.

Backbone Network. The underlying architecture is based on the ResNet-50 (He et al. 2016), known for good classification performance on face and voice data. The fully-connected layer at the top of the original network was replaced by three layers in the following order: a flatten layer, a 512-dimensional fully-connected layer whose output represents the embed-



a)

b)

Figure 3: **Experimental Evaluation Overview**. The steps for training and testing protocols of the face modality (a) and the steps for training and testing protocols of the voice modality (b).

ding features used throughout the experiments; and an output layer whose implementation depends on the loss function integrated in the corresponding network.
Loss Functions. In order to enable the backbone network learns discriminative features, several instances of the network were independently trained through different loss functions from various families. The Softmax loss (Taigman et al. 2014) and its variations, called Center loss (Wen et al. 2016) and Ring loss (Zheng et al. 2018), represented the cross-entropy-based family. Additive Margin loss (Wang et al. 2018a) served the angular-margin-based family.
Training Details. For each possible pair of modality and loss function, a different ResNet-50 backbone network was separately trained on top of the *VoxCeleb Train Split* data. The models were 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; it run on 4 GPUs in parallel.

4.3 Evaluation Protocols

Re-Identification. For each testing dataset, the protocol aims to evaluate how the trained models 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. VoxCeleb Test Split has the minimum number of participants among the considered datasets (i.e., 40). For each user, we 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, the output of the last layer of the ResNet-50 instances was considered as feature vector associated to each frame/spectrogram. 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 Angular Margin 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., Zheng et al. 2013). The probe image is matched against a set of gallery images, obtaining a ranked list according to their matching similarity. The correct match is assigned to one of the top ranks, the top one rank in this case (Rank-1). 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. For each testing dataset, the protocol aims to evaluate how the trained models 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 output of the last layer of the ResNet-50 network instances was 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 (Jain et al. 2000). EER indicates that the proportion of false acceptances is equal to the proportion of false rejections. The lower the EER, the higher the performance. Lastly, starting from the subject selection, the experiment was repeated and the results were averaged.

It should be noted that the above choices allow to evaluate performance on comparable and reasonable numbers of samples among the different datasets.

4.4 Face Evaluation Results

Figure 5 provides the results obtained for both face re-identification and face verification on the selected testing datasets. As it might be expected, the model performance decreases when we move from semi-controlled to uncontrolled scenarios (robot assistance recordings in *AveRobot* VS mobile recordings in MOBIO), while they remain more stable between datasets coming from scenarios which could be comparable in terms of wildness level (mobile recordings in MOBIO VS interview recordings in VoxCeleb Test Split and indoor surveillance recordings in MSU-AVIS VS robot assistance recordings in *AveRobot*). The general system performance degrades due to the lower resolution, bad illumination and pose variations of the captured face images. For face re-identification, *AveRobot* provides inferior performance with respect to the traditional audio-visual datasets. We achieve between 57% and 64% of rank-1 accuracy, more than 10% lower than the performance of the nearest dataset, MSU-AVIS. For verification, the margin over the two datasets is narrower, but there is still a significant decreases in performance with respect to MOBIO and VoxCeleb Test Split. The results confirm that Softmax is not sufficient to train discriminative features. In fact, it is outperformed by the models trained with other loss functions. Overall, the results obtained on VoxCeleb Test Split and MOBIO are better in comparison to the ones observed for *AveRobot* and MSU-AVIS. The experiments highlight the need of more advanced algorithms capable of mitigating the impact of the challenging conditions on the performance.

4.5 Voice Evaluation Results

The results obtained for both voice re-identification and voice verification are depicted in Figure 5. The tasks are challenging since we consider spectrograms obtained by one second of speech and we compute the results based on the comparison of such short

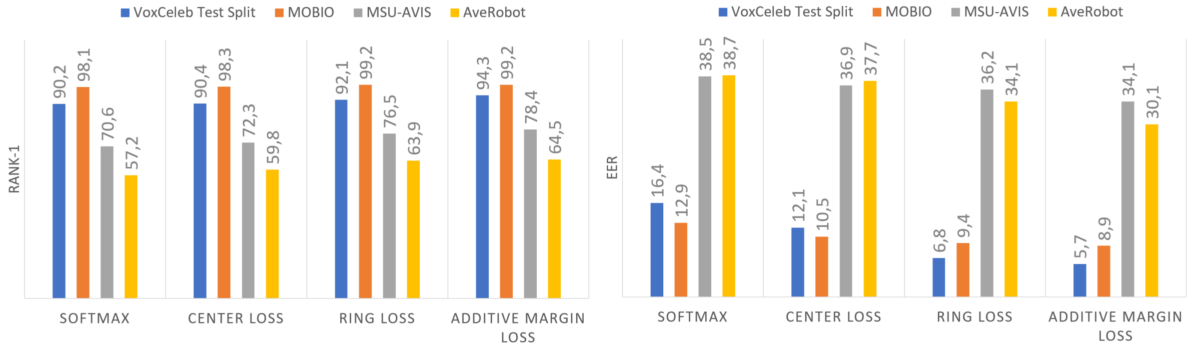


Figure 4: The results obtained by ResNet-50 trained with various loss functions on VoxCeleb Train Split and tested on unseen users from different datasets for face re-identification through Rank-1 (**left**) and verification through EER (**right**).

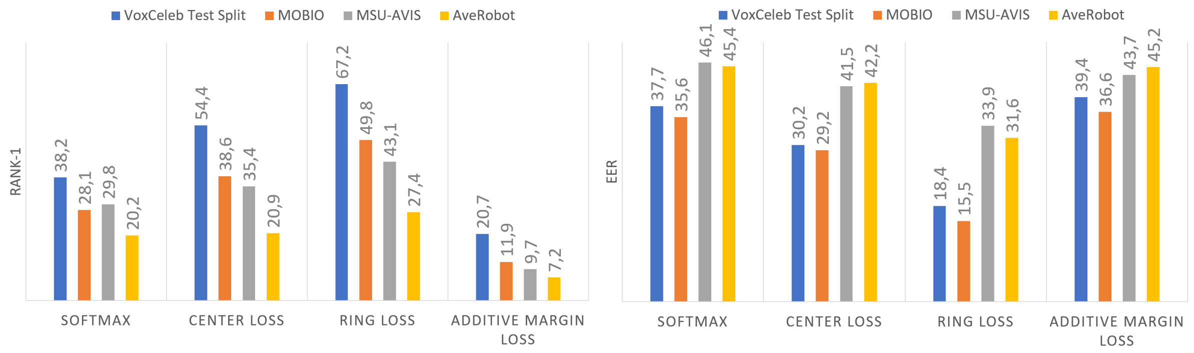


Figure 5: The results obtained by ResNet-50 trained with various loss functions on VoxCeleb Train Split and tested on unheard users from different datasets for voice re-identification through Rank-1 (**left**) and verification through EER (**right**).

spectrograms. The results show that the voice recognition performance is badly affected by the background noise presented in semi/no-controlled scenarios like MSU-AVIS and *AveRobot*. In particular, recognizing people from their voices in *AveRobot* is more challenging in comparison with the other datasets. This observation could derive from the fact that the audios in *AveRobot* contain several noisy situations (e.g., opening doors, background speaking, alarm sounds). The performance improves in more controlled scenarios. Furthermore, for voice re-identification tasks, the gap between *AveRobot* and the other datasets is larger with respect to the face re-identification task. For re-identification, we achieve between 7.3% and 27.4% of rank-1 accuracy. It should be noted that, for voice re-identification, a random guesser reaches 2.5% of rank-1 accuracy. For verification, we get between 31.6% and 45.4% of EER. The Angular Margin loss seems to badly learn the patterns behind spectrograms, while it works well for face images. Overall, the results demonstrate that voice recognition models suffer the most from the challenging recording conditions with respect to face recognition models.

5 CONCLUSIONS

In this paper, we proposed a pipeline for collecting audio-visual data under a multi-floor robot cooperation scenario and leveraged it in order to create a multi-biometric dataset comprising of face and voice modalities, namely *AveRobot*, tailored for evaluating people re-identification and verification capabilities of robots. It includes 111 participants and over 2,500 short videos. In order to establish benchmark performance, different techniques for training deep neural networks on face and spectrogram images, extracted directly from the frames and the raw audios, were tested on this new dataset for re-identification and verification. The performance on this new dataset were compared against the performance the same techniques obtain on other traditional audio-visual datasets from different scenarios. The results demonstrated that *AveRobot* appears as challenging due to the uncontrolled conditions and remarked the need of better understanding how the existing algorithms react against in-the-wild operational contexts.

In the next steps, we plan to explore other deep learning architectures and methodologies to (i) combine sequence of faces/spectrograms coming from the same recording, (ii) merge information coming from faces and voices, (iii) mitigate the impact of the conditions posed by our scenario on face and voice re-identification and verification, and (iv) validate the developed methods on real-world robot assistance.

ACKNOWLEDGEMENTS

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), 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.

Lastly, we would like to thank Nelson González-Machín and Enrique Ramón-Balmaseda for supporting us during the acquisition of the recordings.

REFERENCES

Barra, S., Casanova, A., Fraschini, M., and Nappi, M. (2017). Fusion of physiological measures for multimodal biometric systems. *Multimedia Tools and Applications*, 76(4):4835–4847.

Barra, S., De Marsico, M., Galdi, C., Riccio, D., and Wechsler, H. (2013). Fame: face authentication for mobile encounter. In *Biometric Measurements and Systems for Security and Medical Applications (BIOMS), 2013 IEEE Workshop on*, pages 1–7. IEEE.

Boratto, L., Carta, S., Fenu, G., and Saia, R. (2016). Using neural word embeddings to model user behavior and detect user segments. *Knowledge-Based Systems*, 108:5–14.

Boratto, L., Carta, S., Fenu, G., and Saia, R. (2017). Semantics-aware content-based recommender systems: Design and architecture guidelines. *Neurocomputing*, 254:79–85.

Chowdhury, A., Atoum, Y., Tran, L., Liu, X., and Ross, A. (2018). Msu-avis dataset: Fusing face and voice modalities for biometric recognition in indoor surveillance videos. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3567–3573. IEEE.

Cielniak, G. and Duckett, T. (2003). Person identification by mobile robots in indoor environments. In *Robotic Sensing, 2003. ROSE'03. 1st International Workshop on*, pages 5–pp. IEEE.

Correa, M., Hermosilla, G., Verschae, R., and Ruiz-del Solar, J. (2012). Human detection and identification by robots using thermal and visual information in domestic environments. *Journal of Intelligent & Robotic Systems*, 66(1-2):223–243.

Cosar, S., Coppola, C., Bellotto, N., et al. (2017). Volume-based human re-identification with rgb-d cameras. In *VISIGRAPP (4: VISAPP)*, pages 389–397.

Cruz, C., Sucar, L. E., and Morales, E. F. (2008). Real-time face recognition for human-robot interaction. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE.

Domínguez-Brito, A. C., Cabrera-Gámez, J., Hernández-Sosa, D., Castrillón-Santana, M., Lorenzo-Navarro, J., Isern-González, J., Guerra-Artal, C., Pérez-Pérez, I., Falcón-Martel, A., Hernández-Tejera, M., and Méndez-Rodríguez, J. (2001). Eldi: An agent based museum robot. In *ServiceRob'2001, European Workshop on Service and Humanoid Robots, Santorini, Greece*.

Faber, F., Bennewitz, M., Eppner, C., Görög, A., Gonsionr, C., Joho, D., Schreiber, M., and Behnke, S. (2009). The humanoid museum tour guide Robotinho. In *IEEE International Symposium onRobot and Human Interactive Communication (RO-MAN)*, pages 891–896.

Fenu, G. and Marras, M. (2018). Controlling user access to cloud-connected mobile applications by means of biometrics. *IEEE Cloud Computing*, 5(4):47–57.

Fenu, G., Marras, M., and Boratto, L. (2018). A multi-biometric system for continuous student authentication in e-learning platforms. *Pattern Recognition Letters*, 113:83–92.

Fenu, G. and Nitti, M. (2011). Strategies to carry and forward packets in vanet. In *International Conference on Digital Information and Communication Technology and Its Applications*, pages 662–674. Springer.

Hansen, J. H. and Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal processing magazine*, 32(6):74–99.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hristoskova, A., Agüero, C. E., Veloso, M., and De Turck, F. (2012). Personalized guided tour by multiple robots through semantic profile definition and dynamic redistribution of participants. In *Proceedings of the 8th International Cognitive Robotics Workshop at AAAI-12, Toronto, Canada*.

Hu, J., Shen, L., and Sun, G. (2017). Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 7.

Irfan, B., Lyubova, N., Ortiz, M. G., and Belpaeme, T. (2018). Multi-modal open-set person identification in hri.

Jain, A., Hong, L., and Pankanti, S. (2000). Biometric identification. *Communications of the ACM*, 43(2):90–98.

Koide, K. and Miura, J. (2016). Identification of a specific person using color, height, and gait features for a person following robot. *Robotics and Autonomous Systems*, 84:76–87.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Liu, H., Hu, L., and Ma, L. (2017). Online RGB-D person re-identification based on metric model update. *CAAI Transactions on Intelligence Technology*, 2(1):48–55.

López, J., Pérez, D., Santos, M., and Cacho, M. (2013a). Guidebot. A tour guide system based on mobile robots. *International Journal of Advanced Robotic Systems*, 10.

López, J., Pérez, D., Zalama, E., and Gomez-Garcia-Bermejo, J. (2013b). Bellbot - a hotel assistant system using mobile robots. *International Journal of Advanced Robotic Systems*, 10.

Lukic, Y., Vogt, C., Dürr, O., and Stadelmann, T. (2016). Speaker identification and clustering using convolutional neural networks. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), Vietri sul Mare, Italy, 13-16 Sept. 2016*. IEEE.

Mac, T. T., Copot, C., Tran, D. T., and De Keyser, R. (2017). A hierarchical global path planning approach for mobile robots based on multi-objective particle swarm optimization. *Applied Soft Computing*, 59:68–76.

Martinson, E. and Lawson, W. (2011). Learning speaker recognition models through human-robot interaction. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 3915–3920. IEEE.

McCool, C., Marcel, S., Hadid, A., Pietikäinen, M., Matejka, P., Cernockỳ, J., Poh, N., Kittler, J., Larcher, A., Levy, C., et al. (2012). Bi-modal person recognition on a mobile phone: using mobile phone data. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, pages 635–640. IEEE.

Munaro, M., Fossati, A., Basso, A., Menegatti, E., and Van Gool, L. (2014). One-shot person re-identification with a consumer depth camera. In *Person Re-Identification*, pages 161–181. Springer.

Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Ouellet, S., Grondin, F., Leconte, F., and Michaud, F. (2014). Multimodal biometric identification system for mobile robots combining human metrology to face recognition and speaker identification. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 323–328. IEEE.

Rosenthal, S., Biswas, J., and Veloso, M. (2010). An effective personal mobile robot agent through symbiotic human-robot interaction. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 915–922. International Foundation for Autonomous Agents and Multiagent Systems.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Shiomi, M., Kanda, T., Ishiguro, H., and Hagita, N. (2007). Interactive humanoid robots for a science museum. *IEEE Intelligent systems*, 22(2):25–32.

Sinha, A., Chakravarty, K., and Bhowmick, B. (2013). Person identification using skeleton information from kinect. In *Proc. Intl. Conf. on Advances in Computer-Human Interactions*, pages 101–108.

Sun, Y., Liang, D., Wang, X., and Tang, X. (2015). Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*.

Susperregi, L., Fernandez, I., Fernandez, A., Fernandez, S., Maurtua, I., and de Vallejo, I. L. (2012). Interacting with a robot: a guide robot understanding natural language instructions. In *Ubiquitous Computing and Ambient Intelligence*, pages 185–192. Springer.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708.

Thrun, S., Bennewitz, M., Burgard, W., Cremers, A. B., Dellaert, F., Fox, D., Hahnel, D., Rosenberg, C., Roy, N., Schulte, J., et al. (1999). Minerva: A second-generation museum tour-guide robot. In *Robotics and automation, 1999. Proceedings. 1999 IEEE international conference on*, volume 3. IEEE.

Trahanias, P., Burgard, W., Argyros, A., Hähnel, D., Baltzakis, H., Pfaff, P., and Stachniss, C. (2010). TOURBOT and WebFAIR: Web-operated mobile robots for tele-presence in populated exhibitions. *IEEE Robotics and Automation Magazine*, 12(2):77–89.

Troniak, D., Sattar, J., Gupta, A., Little, J. J., Chan, W., Calisgan, E., Croft, E., and Van der Loos, M. (2013). Charlie rides the elevator–integrating vision, navigation and manipulation towards multi-floor robot locomotion. In *Computer and Robot Vision (CRV), 2013 International Conference on*, pages 1–8. IEEE.

Wang, F., Cheng, J., Liu, W., and Liu, H. (2018a). Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

Wang, Y., Shen, J., Petridis, S., and Pantic, M. (2018b). A real-time and unsupervised face re-identification system for human-robot interaction. *Pattern Recognition Letters*.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Zheng, W.-S., Gong, S., and Xiang, T. (2013). Reidentification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):653–668.

Zheng, Y., Pal, D. K., and Savvides, M. (2018). Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5089–5097.