Audio-video biometric recognition for non-collaborative access granting

Christian Micheloni, Sergio Canazza, Gian Luca Foresti [a,b,a]

[a] *Department of Computer Science, University of Udine, Via delle Scienze 206, 33100 Udine, Italy*

[b] *Department of Historical and Documentary Sciences, University of Udine, Via Petracco 8, 33100 Udine, Italy*

Key words: Face Recognition, Face Detection, Audio de-noising, Speaker recognition

Abstract

In this paper, the problem of non collaborative person identification for a secure access to facilities is addressed. The proposed solution adopts a face and a speaker recognition techniques. The integration of these two methods allows to improve the performance with respect to the two classifiers.

In non collaborative scenarios, the problem of face recognition first requires to detect the face pattern then to recognise it even when in non frontal poses. In the current work, a histogram normalization, a boosting technique and a linear discrimination analysis have been exploited to solve typical problems like illumination variability, occlusions, pose variation, etc. In addition, a new temporal classification is proposed to improve the robustness of the frame-by-

frame classification. This allows to project known classification techniques for still image recognition into a multi frame context where the image capture allows dynamics in the environment.

For the audio, a method for the automatic speaker identification in noisy environments is presented. In particular, we propose an optimization of a speech de-noising algorithm to optimise the performance of the Extended Kalman Filter (EKF). To provide a baseline system for the integration with our proposed speech de-noising algorithm, we use a conventional speaker recognition system, based on Gaussian mixture models and Mel Frequency Cepstral Coefficients (MFCCs) as features.

To confirm the effectiveness of our methods, we performed video and speaker recognition tasks first separately then integrating the results. In particular, two different corpora have been used: a) a public corpus (ELDSR for audio and FERRET for images) and b) a dedicated audio/video corpus, in which the speakers read a list of sentences wearing a scarf or a full-face motorcycle helmet. Experimental results show that our methods are able to reduce significantly the classification error rate.

1 Introduction

Security is getting everyday more importance in the research field. Providing protection by means of new technologies is very stimulating for the research community. Within the security domain, assessing people identity is certainly one of the most challenging problems.

Biometrics have been recently referred to the study of methods for people

recognition on the basis of their characteristics. Two main traits can be exploited for such a purpose:

- physiological: consists in physical properties of the body like shape, structure, etc. These traits are those that the person cannot change or train.
- behavioural: consists in behavioural properties like walking posture, keystroke, voice, etc. These traits are those that the person learns in doing a task.

It is therefore possible to recognise a person by analysing its physical properties or the way to perform an operation. On this point of view, we can see how the former can be usually elaborated working on a single time sample while the latter normally requires a temporal analysis.

Although there are differences in the two approaches, both select human traits on the basis of the following parameters [1]:

- Universality: each person should have the characteristic;
- Distinctiveness: any two persons should be sufficiently different in terms of the characteristic;
- Permanence: the characteristic should be sufficiently invariant (with respect to the matching criterion) over a period of time;
- Collectability: the characteristic can be measured quantitatively.

By simply following these parameters, a really mature technology for person identification is represented by fingerprint and DNA recognition. Although these technologies have even proven forensic capability, they do not fulfill a fundamental requirement for large scale recognition: *acceptability*. This is a real important factor for determining the success of a technology. Indeed, a person is not willing to press the thumb on a fingerprint reader or to give a

DNA sample for getting access to a restricted zone. Therefore, providing a technology that is as less invasive as possible is of highest importance.

Under such considerations, the visual aspect of a person, acquirable with a video camera, and the characteristic way everybody talks, are definitively two technologies that deserve further investigation to propose them as mature technologies as those given by fingerprint or DNA.

With respect to visual recognition, there exist different techniques for analysing both physiological and behavioural traits. In this work, we want to address only those considering the physiological characteristics by focusing on techniques for face recognition. Such a technology is well accepted by citizens since it could be adopted without requiring to touch a sensor (i.e. fingerprint) or to be beamed by active sensors (i.e. retina).

The majority of the available works deals with the problem of face recognition by assuming that the person is collaborative with the system. The face recognition task is considered as a stand alone pattern classification problem and more specifically as a single still image classification problem. Such assumptions are not valid in a non collaborative scenario. In this domain, people are not limited in their movements thus increasing the problem complexity. In particular, the task to detect the face is mandatory: it can bring into the following recognition step additional errors with respect to the case in which still images are used. Moreover, the possibility of detecting faces even when they are in non frontal poses stresses even more the algorithms that well perform on still images.

A lot of methods have been proposed to address the problem of face detection, e.g., Support Vector Machines (SVMs) [2], Neural Networks (NNs) [3], Hidden

Markov Models (HMMs) [4] or boosting algorithms [5]. For an exhaustive survey, the reader is referred to the paper of Yang et al. [6].

Within the past two decades, several face recognition methods have been developed. In [7] a detailed survey classifies the existing approaches in three general categories: holistic, feature-based and hybrid techniques. Anyhow, many current methods do not consider important factors like the pattern quality (resulting from the detection) and the grade of possible occlusions. Indeed, many works show results obtained on important face datasets acquired in controlled environments or with ground-truth markers for image cropping and alignment [8, 9].

Few methods consider the problem of face recognition in a non cooperative scenario [10, 11]. These methods try to extend existing algorithms developed for still images (single sequence frames), by applying probabilistic or majority voting mechanisms to accumulate the recognition scores obtained on single consecutive frames. However, fundamental spatio-temporal information is not taken into account. Zhou et al. [11] have proposed a probabilistic algorithm for applying still-to-video and a video-to-video face tracking and recognition. The algorithm is based a Particle Filter (PR) method specifically modified to increase the computational efficiency.

In this paper, the problem of non collaborative person identification for a secure access to facilities is addressed. The main objective is to obtain a natural and non-intrusive interaction with the authentication system. This specificity is very important in the adopted application domain that requires a technology that has to be made as transparent and easy to use as possible.

In the proposed work, to increase the robustness of the system, the face recog-

nition module is integrated with a speaker recognition module. Speech is the result of a combination of physical traits (important for speech signal segmental features, such as voice formant positions) and behavioural patterns (important for supra-segmental features, such as prosody). Behavioural patterns in the sense that speaking is something we do. Children grow learning to speak in an environment influenced by many social and linguistic factors (language, dialect, social status, etc.): this environment influences how we speak. The speech is also affected by physical characteristics (nasal cavities, vocal folds size and shape, vocal tract), and these factors influence how we speak and how the speech sounds. In a non-collaborative scenario (where the speaker's actions are not finalized to his/her identification), specialized measurement devices (electropalatographs, electroglottographs, etc.) are inaccessible. Therefore, all information from the speech process must be extracted by sampling the speech signal captured through one or more microphone devices.

In this paper, the speaker identification in a non-collaborative single speaker scenario is addressed. The adopted method is text-independent and user-driven: users are allowed to say anything during both enrolment and test phases. Non-collaborative scenario is often characterized by noisy environments (i.e. audio signal has a low Signal-to-Noise Ratio, SNR). To improve noise robustness of speaker identification we propose a time domain algorithm based on the Extended Kalman Filter (EKF) optimized for speech de-noising.

Experimental results with non-collaborative subjects (someone wears a full face helmet, others a scarf, etc.) show that the proposed method is able to significantly reduce the error rate (false and missed alarms) and to reach a satisfactory level with respect to existing identification methods working with collaborative persons.

2 Face Recognition

Face recognition is a really challenging task. Moreover, if we relax the constraint about the cooperativeness of the individual and the controllability of the acquisition environment the challenge becomes even more demanding.

As matter of fact, in a non cooperative mode and in a free environment, we have to deal with the following factors [12]:

- Illumination - state of the art algorithms perform well on standard datasets whose test images have been acquired almost in the same conditions of the training images. In real security applications, surveillance cameras [13] adapt their intrinsic camera parameters on the scene to obtain the best quality. This implies that, due to the illumination changes caused by different factors (e.g. , environment reflectance, sunlight, etc.), the appearance of a subject could be noticeably different between two different time instants. To overcome such a problem, different techniques can be adopted to normalize images.
- Pose - in a non cooperative mode we have to deal with different acquisition poses that generate quite different patterns with respect to the training ones. In these conditions, the performances of the algorithms suffer a considerable reduction if compared to those obtained on standard datasets.
- Occlusions - working in a non controlled environment, people can move without restrictions, thus each person can occlude any other. In addition, we can have situations where occlusions are introduced by the clothes or accessories a person is wearing (e.g., scarf, helmet, etc.).

A fourth problem described by Abate *et al.* [12] is represented by the *time*

delay. This causes a non linear transformation of the face patterns due to the aging process in which every person is involved. In this work, we do not consider such a problem as it can be partially solved by periodically updating the watch-list then training the classifiers.

Instead, for the other three problems we propose a step-by-step procedure. Once the image is acquired we perform a histogram equalization to normalize the image contrast in both training and test images. Then, a face detection algorithm is executed to detect the face and/or part of it. In particular, the AsymBoost* algorithm [14] has been used to train multi-layer cascade classifiers able to detect multiple pose face patterns. The detector is based on the coarse-to-fine strategy introduced in [15] and it is similar to the cascade concept. The first levels have been trained to detect a generic human face shape, while the subsequent levels detect a more specific pose. In particular, the proposed technique deals with out–of–plane rotations, in the range $\Theta = [-90, +90]$ degrees with a 15 degrees sampling. This subdivision corresponds to the different levels and to different classifiers. At every level, a different classifier has been trained with patterns belonging to the correspondent view range. Each classifier of this multi-layer detector corresponds to a cascade of Haar based weak classifiers trained by AsymBoost* algorithm. This technique allows to detect face pattern in different poses but with limited degree of occlusion. To sidestep such a problem we trained, by using the same algorithm, a cascade of classifiers for detecting the eyes of the person. The detection of such features has a double use depending on whether the entire face has been detected or not. In the first case, the detection of the eyes is fundamental to apply a transformation for the alignment of the face pattern with respect to the ones present in the database. In the second case, from the

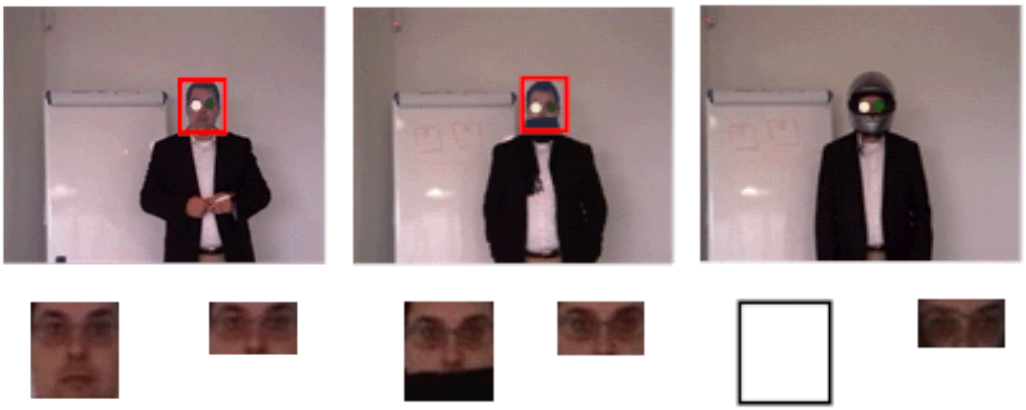

Figure 1. *Example of the face detection phase on three different cases. (a) In the normal situation both the entire face and the eyes are detected. In addition, the eyes positions are exploited to crop the original detected pattern and to rotate it in order to align the pattern with respect to those in the database. (b) When the face is slightly occluded the face detector is still able to return a pattern. Though, this can produce bad results during the recognition process. For this reason the eyes positions are used to determine a reduced section of the face to be utilized for the recognition phase. (c) In case of really occluded face, the detection of the whole face fails. Hence, we adopted the results of the eyes detection for cropping an area surrounding the eyes for the recognition process.*

position of the eyes, a wider region is extracted for the recognition process. The complete detection process is summarised in Fig. 1.

Once a face pattern or a part of it is detected, a linear discrimination analysis is adopted to project the face patterns into a space where the linear separability between the classes is optimized. Let $S = \{x_1, \ldots, x_n\}$ be a set of p-dimensional samples belonging to c different classes c_i of $d_i, i = 1, \ldots, c$ dimension each. The linear discrimination analysis seeks a linear projection w that maximizes the between class scatter and minimizes the within scatter of

the projected samples by maximizing:

$$J(w) = \frac{w^T \mathbf{S}_B w}{w^T \mathbf{S}_W w} \tag{1}$$

where the between class scatter matrix $\mathbf{S}_B$ is given by

$$\mathbf{S}_B = \sum_{i=1}^{c} d_i (\mu_i - \mu)(\mu_i - \mu)^T \tag{2}$$

where μ_i and μ are respectively the p-dimensional means of the i-th class and of the entire set S. While the within class scatter matrix is given by

$$\mathbf{S}_W = \sum_{i=1}^{c} \sum_{x \in c_i} (x - \mu_i)(x - \mu_i)^T \tag{3}$$

To maximise (1) the following eigenvalue problem has to be solved:

$$\mathbf{S}_W^{-1} \mathbf{S}_B w = \lambda w \tag{4}$$

As can be noted, such an equation has two main problems. The first consists to the fact that $\mathbf{S}_W$ has to be non-singular in order to compute its inverse. If the number of classes c is lower than the dimensionality p of the samples, $\mathbf{S}_W$ is singular. If we consider that face pattern sizes are usually lower-bounded by a size of 25×25 we are required to have at least 625 different classes to get a non singular $\mathbf{S}_W$ matrix. Instead, since we are working on what is commonly called a Small Sample Size (SSS) problem, we need a technique that reduces the dimensionality of the original patterns. For this reason it is common to adopt a pre-processing step in which a PCA compression is applied to the original data.

The second problem in (4) is represented by the rank of $\mathbf{S}_B$ that being at most $c - 1$ limits the number of non-zero eigenvalues to $c - 1$. This is a stringent limitation especially in those cases where the number of classes is really small. As matter of fact, in the selected watch-list case we could have a very small

number of classes. To overcome such a problem Xiang *et al.* [9] proposed a recursive solution. This, after extracting a feature w_j discards its information from all the sample vectors x_i using the followin process:

$$x_i^{j+1} = x_i^j - \left(w_j^T x_i^j\right) w_j \tag{5}$$

where $x_i^1 = x_i$. Since this process is computationally demanding, the authors proposed a faster method to compute the $j-th$ scatter matrices from the previous ones rather then rebuilding the entire training set as in (5). For further details the reader is referred to [9]. In the current work, the regularized LDA (R-LDA) [16] and the recursive FLD (RFLD) [9] have been adopted to train the classifiers. Once each of the two method extracted the first k features, the training patters have been projected into the new space for computing the covariance matrices Σ_i for each $i-th$ class. Such an information is used to compute the distance of a projected test pattern $\hat{y}$ from a class i by using the Mahalanobis distance that is:

$$D_M^i(\hat{y}) = \sqrt{(\hat{y} - \hat{\mu}_i)^T \Sigma_i^{-1} (\hat{y} - \hat{\mu}_i)} \tag{6}$$

where $\hat{\mu}_i$ is the mean of the projected training patterns of the $i-th$ class.

The distance is then normalised by

$$\bar{D}_M^i = \frac{D_M^i}{\sum_{j=1}^{c} D_M^j} \tag{7}$$

To associate a higher classification probability to the closer classes we have defined the following rule:

$$P(x \in c_i) = \frac{1 - \bar{D}_M^i}{c - 1} \tag{8}$$

Finally, the classification is given by returning the class whose probability is

the highest. It is worth noticing that when the detector returns occluded patterns (e.g. scarf samples) the recognition probability drops down dramatically. In such cases, we consider failed the recognition and we try with the region of the eyes if available.

3 Speaker Recognition

The speaker recognition task is usually divided into speaker verification and speaker identification.

a. Speaker verification is finalized to determine if an identity claim is true or false.
b. Speaker identification is a $N + 1$-class problem finalized to determine if the current speaker is a known speaker and to identify he/she among the N known target speakers.

In the second case further subtask can be identified:

- Speaker change detection - to detect different speakers in a conversation.
- Speaker tracking - to track a given target speaker during a conversation.
- Speaker clustering - to group *similar* speakers accordingly to a similarity measure.
- Speaker diarization - to assign a label to every speaker in a conversation and to group the speeches of the same speaker.

In order to recognize the speech of a known target speaker, the system should perform the speaker enrolment by processing an audio sample of speaker's speech.

This work deals with the speaker identification in a non-collaborative single-speaker domain. The system used is text-independent and user-driven [17]: a fully text-independent system. Users can say anything during the enrolment (in this experiment we used the same sentence for different languages) and test phases. Non-collaborative scenario is often characterized by noisy environments (i.e. audio signal has a low Signal-to-Noise Ratio, SNR). To improve the robustness to noise for the speaker identification, we propose a time domain algorithm based on the Extended Kalman Filter (EKF). Experimental results with non-collaborative subjects (wearing a scarf or a full face helmet) demonstrated that our method is able to significantly reduce the error rate. A conventional speaker recognition system has been adopted as baseline for the proposed speech de-noising algorithm. Such a classifier is based on Gaussian mixture models [18] and Mel Frequency Cepstral Coefficients (MFCCs) as features.

3.1 Speech de-noising

To improve the robustness to noise for the speaker identification, the most popular methods have been considered.

The de-noising algorithms can be divided into three categories: 1) *frequency-domain* methods, such as different forms of non casual Wiener filtering or spectral subtraction schemes [19, 20] and recent algorithms that attempt to incorporate knowledge of the human auditory system [21, 22]; these methods use little a priori information (only the Power Spectral Density noise estimation); 2) *time-domain* restoration by signal models such as Extended Kalman filtering [23–25]: in these methods, to estimate the statistical description of

the audio events (speech vs. music, for example), a considerable amount of a-priori information is necessary; 3) restoration by *source models*: to develop the physical model of the source, only a-priori information [26] is used.

The advantage of frequency-domain methods is that they are straightforward and easy to implement. However, the limitations are : a) musical noise (short sinusoids randomly distributed over time and frequency) is unavoidable; b) the results depend on a good noise (and SNR) estimation; c) the power spectrum of the background noise has to be known in advance. Restoration by source model is very limited to few cases and cannot be generalized. To solve this problem, we propose a time domain algorithm, optimized for speech denoising, that uses the Extended Kalman Filter theory (EKF) in the implementation proposed by M. Niedźwiecki and K. Cisowski [23, 27]. We observe that the algorithm in [27] can be interpreted as the nonlinear combination of two Kalman filters: the first is used to follow the slow variations of the signal time-varying AR model parameters, while the second takes part in the reduction of background and impulsive noise. At medium and high Signal-to-Noise Ratios (say, SNR $\geq 10\,\mathrm{dB}$), the performance of such a filter is superior to that of other standard methods like spectral attenuation. Anyhow, a simple use of such a technique does not guarantee the best results. This is due to the non-stationary characteristic of audio signals that yields to errors in parameter tracking and noise filtering, especially during fast transients. It means that, in order to achieve maximum performance from the EKF, an optimization of its implementation is mandatory. In this work, we propose to improve the algorithm in [27] in order to deal with the non-stationary nature of the audio signal. In particular, we use a more performing model tracking procedure and a refined bootstrapping strategy. The careful combination of the proposed

techniques and an accurate choice of some critical parameters allows to improve the performance of the EKF algorithm. We implemented the algorithm as a plug-in based software tool (using Microsoft DirectX technology), which can be used as an added module to most used audio software. *Real time* computation (with latency time of $200ms$, approximately) is reached on a system based on a 3.2 GHz Intel Xeon quad-core with Windows XP.

3.2 Problem statement

Let the audio signal $s(t), t = 1, 2, ...,$ be modelled by a *time varying* autoregressive (AR) model of order p

$$s(t+1) = \sum_{i=1}^{p} a_i(t)s(t-i+1) + e(t) \tag{9}$$

driven by the Gaussian zero-mean white noise sequence $e(t)$ with variance σ_e^2. The evolution of the time varying coefficients $a_i(t)$ is modelled by the random walk model

$$a_i(t+1) = a_i(t) + w_i(t), \quad i = 1, ..., p \tag{10}$$

with $w_i(t)$ zero mean Gaussian white processes of variance σ_w^2 mutually uncorrelated, i.e., $\mathrm{E}[w_i(t)w_j(t)] = 0$ for $i \neq j$, and independent of $e(t)$.

Moreover, let us assume that the original signal $s(t)$ is corrupted by a mixture of a broadband noise (environmental noise) $z(t)$ and impulsive noise (artefacts in analog/digital audio recordings or - in general - a real signal characterized by a short duration, a random occurrence and a high power spectral density) $v(t)$ (independent of $e(t)$ and $w_i(t)$), so that the available signal $y(t)$ can be written as

$$y(t) = s(t) + z(t) + v(t) \tag{11}$$

The noise $z(t)$ is assumed to be Gaussian zero-mean white noise (in the case of a coloured noise $z(t)$, based on a estimated environment noise, it suffices to model it as an AR process and to increase the state dimension accordingly [28]) with variance σ_z^2, while $v(t)$ is assumed Gaussian zero-mean noise with $\sigma_v^2 = \infty$ if a click is present, or $\sigma_v^2 = 0$, otherwise. As a consequence, if a click is revealed at time t, the corresponding sample $y(t)$ must be discarded since it does not bear information on $s(t)$ and $s(t)$ must be recovered from $\{..., y(t-1), y(t+1), ...\}$.

In [23], it is shown that under the above hypotheses, the problem of recovering the signal $s(t)$ from the noisy measurements $\mathbf{Y}(t) = \{y(t), y(t-1), ..., y(1)\}$ can be optimally handled by an Extended Kalman Filter (EKF).

3.3 Improvements

Bootstrap procedure The first problem we deal with, is the choice of the filter initial conditions. For such a purpose, let us notice that starting the algorithm from scratch implies an initial transiency of the parameter tracker during which the EKF noise reduction capabilities are greatly reduced. To solve this problem, we introduced a bootstrap procedure: the first t_0 ms of the signal are time-reversed and fed to the filter (we adopted $t_0 = 50ms$. In this way, the parameters for a proper initialization of the model are estimated. Hence the restoration of the "true" signal will use such values as the initial conditions.

Stability check The estimated time-varying AR-model is not guaranteed to be stable at all times but, in practice, instability is quite a rare event. How-

ever, it may severely interfere with proper click detection, we preferred to test stability at each time t via the Levinson recursion [29], i.e. testing if the magnitude of the corresponding reflection coefficients is less than the unity (the computational overload is negligible). In the case of an unstable AR-model, the parameter update is skipped.

Variable forgetting factor A very delicate part of the filter (in particular, when it is used as audio restoration tool) is the tracking part. By exploiting a variable memory we can have "long memory" in almost-stationary cases (allowing a better smoothing of coefficients) and "short memory" (fast reaction) in model transients. Such adaptive tracking is also known as "Variable Forgetting Factor" (VFF) [30]. The main idea behind it is to set the memory length inversely proportional to the average information of the acquired samples. For example, a time varying fourth-order (two formants) synthetic AR process with Gaussian noise added (SNR $= 20$ dB) was applied to the EKF filter. Fig. 2 presents the coefficients' true trajectories (dashed line) and the ones estimated by the VFF (continuous line) and the Exponentially Weighted Least Squares (EWLS [30]) by means of dash dotted line. The transition is linear between two stationary models and it is clear that VFF algorithm response is faster than EWLS without increasing the variance.

The speech signals in our audio/video corpus have an average SNR $= 20$ dB. The signals are segmented into frames up to $200ms$ long and each frame is de-noised using both EWLS (with parameter $\gamma = 1 - 10^{-5}$) and VFF ($\gamma_{\min} = 1 - 10^{-4}$) algorithms, in order to test their performances. For the last one, several values of the forgetting factor were used (with the notation in [30], $q_0 \in [0.1 \div 10.0]$); it is important to notice that a smaller value could lead to several interventions of stability check. We verified that VFF algorithm has

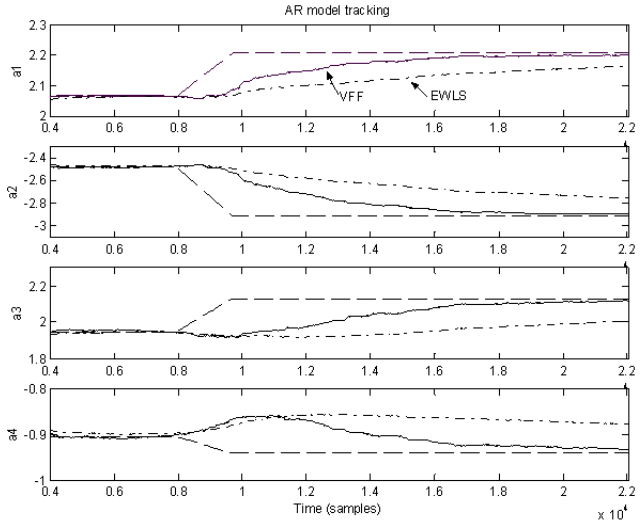

Figure 2. *AR coefficients (*$\mathrm{SNR}_i = 20$ dB*); true trajectories (dashed line), estimated ones by the VFF (continuous line) and the EWLS (dash dotted line); particular of the transition zone.*

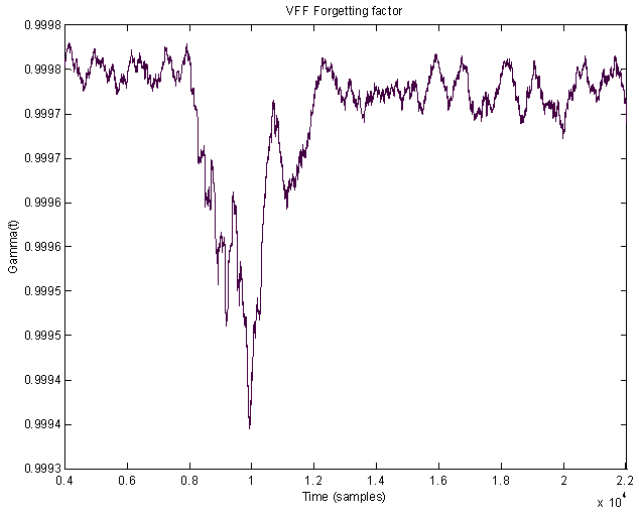

Figure 3. *VFF forgetting factor of Fig. 2*

a better performance in the high frequency zone than EWLS: however these differences don't improve the result of the speaker recognition process.

Clicks detection uses the procedure explained in detail in [31]. This tool is not been improved, because the low click-rate present in the considered recordings.

In this work Mel Frequency Cepstral Coefficients (MFCC) [32] are used. The centre frequencies of band pass filters are equally spaced on a mel scale, and cepstrum vectors are computed from filter log amplitudes through a cosine transform. In the computation of MFCC vectors (such as the number and shape of filters and the number of cepstral coefficients), the parameters setup used was that one standardized by ETSI (ETSI ES 201 108, http://www.etsi.org). This setup was tested by [33] on the YOHO corpus with good results. The use of a filter bank and its frequency spacing according to the mel scale in the MFCC can be motivated as approximations of a basic psychophysical function in the human auditory system, namely the frequency resolution in the cochlea (critical bands).

The choice of the classification method depends on the application of interest [34]. Vector Quantization (VQ) is often used in text-independent scenario [35, 36]: a codebook is trained for each target speaker by optimizing the location of the centroid vectors relative to the target's enrolment speech. The test metric is the average distortion introduced when using a target's codebook to code a test utterance: the smaller distortion, the more similarity between enrolment and test speech.

The Hidden Markov Models (HMM), can be text-dependent (left-right HMMs [37]) or text-independent (ergodic HMMs [38]). HMMs are functional to model spectral dynamics, because they use statistical models instead of template vectors to represent observation vectors in training data. HMMs are trained on repetitions of a chosen unit (sub-word, word or phrase) by the target speaker.

During training, the parameters of the HMM are chosen to optimize some criterion. The test metric is usually the likelihood that the model generated.

The Gaussian Mixture Models (GMMs) are is inherently text-independent [39, 40]. The parameters of a (target) GMM are trained to optimize some criterion defined on enrolment data from a target speaker, and the test metric is the likelihood that a model generated some observed test data. We use 512-component GMM, jointly trained for male and female speakers. No score normalizations (such as T- or Z-norm) [41] are performed.

4 Audio-Video Recognition

In the proposed solution, we have investigated the possibility of integrating two different recognition methods in order to achieve a more robust result. In particular, we want to fuse audio with video information giving a recognition that comes from both physiological and behavioural traits. Since both classifiers return the probability that an input pattern belongs to a class inside the watch-list, we could think that these two processes are independent. This allows to determine the association probability as:

$$P(x \in c_i) = P_A(x \in c_i) \cdot P_V(x \in c_i) \tag{12}$$

where P_A and P_V are respectively the probabilities returned by the speaker and video classifiers.

The main problem in using such a solution is given by the fact that while the speaker classifier needs a time interval for the feature extraction and therefore to give a probability, the video classifier returns a probability for

each frame (every $1/25s$). Hence, to overcome such a problem, we studied a triggering mechanism that is summarized by the following pseudo-code:

repeat

collect video probabilities $P_V^t(x \in c_i)$

until $t > TI$

Extract audio features in the time interval TI

Compute $P_A(x \in c_i) i = 1, \ldots c$

Compute $P_V^{TI}(x \in c_i) i = 1, \ldots c$

Compute $P(x \in c_i) = P_A(x \in c_i) \cdot P_V^{TI}(x \in c_i)$

where $P_V^{TI}(x \in c_i)$ is defined by:

$$P_V^{TI}(x \in c_i) = \prod_{t \in TI} P_V^t(x \in c_i) \tag{13}$$

considering independent the probabilities that the patterns belong to a class c_i at different time instants.

With regard to the speaker recognition, all the parameters used in the speech de-noising (in particular: γ, p, μ, σ_z and the starting value of σ_e^2 [27,30,31,42]) are set in order to reach the best algorithm performance, without particular respect to the real time: as matter of fact, to fusing audio with video information allows to subdivide the audio track in different segments ($2200ms$). In this sense:

- the parameter γ must be chosen in accordance with the non-stationary degree of the signal. We determine that a constant value $\gamma = 1 - 10^{-5}$ is adequate in the case of speech signal;
- for parameter μ, a small value (~ 4) allows the detection of small clicks but

introduces many false detections. We decided to use a higher μ value (i.e. $\mu = 6$);

- the m and q values depend on the particular signal: considering $p = 12$, we chose $m = 0.8$ and $q = 12$ ($q \geq p$ is the signal vector, using the formalism defined in [30, 31]) that provide an average SNR gain of about 10 dB for $\text{SNR}_I \simeq 10$ dB as the minimum average SNR_I value for which the algorithm gives satisfactory results.

The SNR_O of the output signal produced by the smoothing algorithm was measured and related to the SNR_I of input signal. In particular the relation described by

$$\text{SNR}_O \simeq m \cdot \text{SNR}_I + q \ \ [\text{dB}] \tag{14}$$

adapts well to $10 \text{ dB} \leq \text{SNR}_I \leq 40$ dB.

The input signal (48 kHz sampling rate) is pre-emphasized, de-noised and divided into $2200ms$ frames with an overlapping period of $1100s$. A Hanning window is applied to each frame. 13-element MFCC vectors are then computed for each frame and delta and double-delta coefficients are appended.

5 Experimental Results

To validate the proposed solution an incremental test phase has been studied to show the effectiveness of the stand alone face and speaker recognition on real non cooperative situations then to show how these two techniques can be complementary when one of the two gives bad results or even fails. In particular, we acquired different videos and audio tracks of people of the Avires Lab.

As matter of fact, many speaker verification corpora exist, covering many languages and different communication channels. Some of them are available by means of LDC (http://www.ldc.upenn.edu/) and ELRA (http://www.elra.info/). In [34], Melin gives complete overview of the existing corpora. Depending on the purpose of the experiments, there may exists a suitable public corpus, or a dedicated corpus has to be collected. In this work, two corpora were used:

a. English Language Speech Database for Speaker Recognition (ELSDSR) [43]. Data Type: speech Data. Source: MARANTZ PMD670 portable solid state recorder. Languages: English. Intersession interval: none. Channels: Wideband. 16 kHz, 16 bits. 22 speakers (12 M and 10 F), and the age covered from 24 to 63. The training text is the same for every speaker in the database. The text was made with the attempt to capture all the possible pronunciation of English language including the vowels, consonants and diphthongs, etc. Seven paragraphs of text were constructed and collected, which contains 11 sentences. For the training set, 154 (7*22) utterances were recorded; and for test set, 44 (2*22) utterances were provided. Corpus of read speech has been designed to provide speech data for the development and evaluation of automatic speaker recognition system. ELSDSR corpus design was a joint effort of the faculty, Ph.D. and Master students from department of Informatics and mathematical modelling (IMM) at Technical University of Denmark (DTU). The speech is spoken by 20 Danes, one Icelander and one Canadian.
b. AVIRES Corpus. Authors: Christian Micheloni and Sergio Canazza, 2008; Univeristy of Udine, Italy. Data Type: video and speech Data. Source: digital video camera and microphone. Languages: English, Italian and Indi. Intersession interval: none. Channels: Wideband. 48 kHz, 16 bits.

The recording sessions were conducted using 6 speakers (5 M and 1 F; the age covered from 28 to 43), where each speaker performs the following tasks: a) reads a list of sentences (to be used in the enrolment session); b) gives a lecture in English language; c) gives a lecture in Italian and Indi languages; d) reads a list of sentences putting on a scarf; e) reads a list of sentences putting on a full-face motorcycle helmet. The tasks (d) and (e) are necessary in order to test the system in a non-collaborative scenario. Each speaker performed this set of tasks using the same equipment setups. The sessions were one day long. We don't focus on the changing affected by the human voice: long-term, mainly due to aging [44], and on shorter terms due to other factors such as health, speech effort level and speaking rate, emotional state [45].

Concerning the video corpora, together with the acquisition of the audio AVIRES corpus we have acquired the footages containing people with no occlusions, wearing a scarf or a full helmet. Typical frames of the acquired sequences can be seen in Fig. 4.

5.1 Face Recognition

The face recognition results deeply depend on the performance of the face and eyes detection. It is indeed fundamental that the classifier works on pattern that are as much as similar to those in the database. This means that detecting the face correctly thus the eyes allows to crop the face pattern in a way it is well aligned and similar to the training patterns.

The first set of experiments has been conducted to assess the performance of

Figure 4. *Samples of the test sequences. Top two rows have been used to validate the face recognition techniques. Bottom two rows have been used to test both speaker and face recognition systems when physical occlusions, as those represented by a scarf or a helmet, can significantly modify the audio and video patterns*

the two selected linear discriminant techniques (R-LDA and RFLD). The test evaluates the effects of using just a small area surrounding the eyes instead of the whole pattern. To make such an evaluation comparable, we have adopted the experimental protocol defined in [16]. In particular, for all the subjects in the FERET, 4 images have been randomly selected to be part of the training set, while 2 patterns for each subject have been selected by the remaining images to be part of the test set. In Figure 5(a) this first evaluation is presented. It is worth noticing how the R-FLD technique outperforms the R-LDA accordingly to the results presented in [9]. Even more interesting is to notice the performance degradation in correspondence of the use of the eyes region as pattern for the recognition. Anyhow, the R-FLD technique still shows a maximum error minor than 20%.

To test the performances of the two linear discrimination techniques on a SSS problem we have recorded in cooperative mode 6 videos of the AVIRES members. From each footage, 4 images have been randomly selected to be part

of the AVIRES dataset. In this phase, since the number of people (i.e. classes) belonging to our watch-list is limited to 6, the maximum number of features usable by the R-LDA technique is 5 (see [16]). To sidestep such a limitation, hence to increase the number of classes, we have populated our training set by including patterns belonging to 40 different people randomly selected from the FERET database [8]. In particular, to extract homogeneous patterns, the adopted face and eyes detection technique has been executed to crop patterns belonging either to the AVIRES corpus or to the FERET. Hence, to evaluate the two techniques and to show the error introduced by the non-coperativeness of the subject as well as by the lower resolution of the images acquired by the CCTV camera with respect to the dataset's ones, we have computed the detection error versus the dimension of the projected space. The test set has been determined by acquiring footages different from the training ones. In this new videos the subjects were allowed to turn the head. Occlusions were not present. In Figure 5(b) the experiments results are plotted. It is interesting to notice the effects of the alignment onto the detection rate. It can be noted that without aligning the face patterns on the basis of the eyes positions, the error rate is around the 70%. Such an error would not allow to adopt any of these two discriminant techniques.

To further analyse this first battery of experiments the confusion matrices are presented in Table 1. In particular, the errors dispersions are presented for the AVIRES extended case with face aligned. The values reported, once again, demonstrate how the RFLD technique performs 5% better even in context of real sequences of non cooperative subjects. It is even interesting to notice the error committed by classifying an AVIRES member as a FERET member (IDX column) and the error committed by classifying a person not included

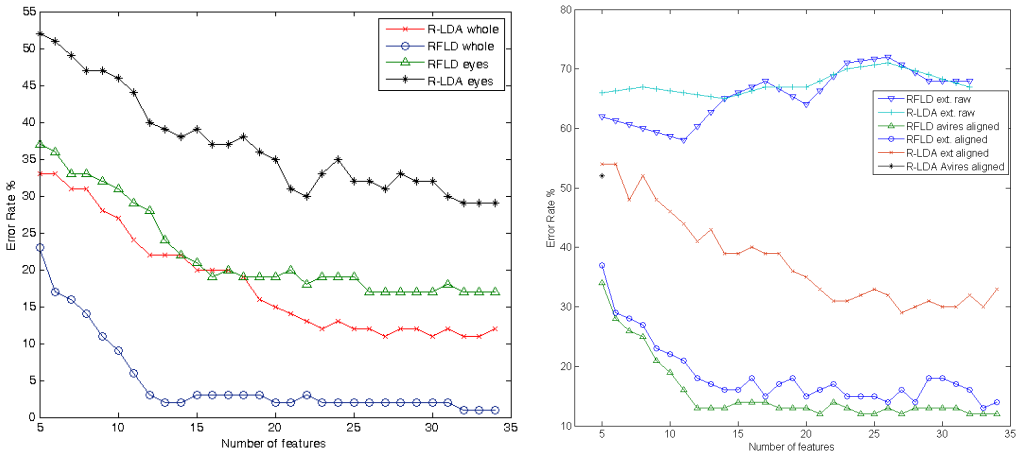

Figure 5. *Evaluation of the linear classifiers. (a) Plots the error rate versus the number of features odopted as dimension of the projecting space. The chart shows the performances of the two selected classifiers with respect to the type of patterns used:whole face or eyes region. (b) Plots the error rate computed on the AVIRES corpus. The R-LDA and RFLD performances are presented on three different cases: i) the AVIRES dataset is extended with FERET patterns and no alignment is executed on the detected faces, ii) the AVIRES dataset is extended with the FERET patterns and the alignment of the face patterns is done on the basis of the eyes detection, iii) the AVIRES dataset is not extended and the face patterns are aligned.*

in the training set (ID7 row).

A second experiment has been conducted to see the effects of using the whole pattern face when some occlusions are present. For such a purpose, we have acquired sequences of the 6 AVIRES members wearing a scarf. As result the entire lower section of the face is occluded. This new condition has not created troubles to the face detector since its performance remained similar to the not-occluded situation. Instead, the recognition process demonstrated a considerable degradation. The values presented in Table 2 show that when the the whole face pattern is passed to the RFLD classifier the correct classification rate falls below the 50%. Instead, if the eyes coordinates are exploited to crop a smaller area around the eyes and such an area is passed to the RFLD

	R-LDA							RFLD						
x \ c	ID1	ID2	ID3	ID4	ID5	ID6	IDX	ID1	ID2	ID3	ID4	ID5	ID6	IDX
ID1	77,33	5,33	4,65	0,00	2,71	2,37	7,61	81,73	3,64	3,47	0,00	2,37	1,69	7,11
ID2	4,82	74,27	4,09	0,00	3,22	1,02	12,57	3,36	81,58	2,63	0,00	1,75	2,49	8,19
ID3	6,00	5,77	76,21	0,00	1,85	2,54	7,62	3,70	3,46	80,83	0,00	1,85	3,00	7,16
ID4	0,20	1,41	1,41	71,98	9,07	6,85	9,07	0,20	1,41	1,41	82,06	3,02	3,63	8,27
ID5	1,84	2,50	2,17	2,17	76,46	5,34	9,52	0,17	0,83	1,17	0,50	83,81	5,01	8,51
ID6	0,00	1,84	1,99	2,15	5,52	77,15	11,35	0,00	0,31	0,92	1,84	3,99	86,35	6,60
ID7	3,33	0,00	3,33	3,33	0,00	3,33	86,67	3,33	0,00	3,33	3,33	0,00	3,33	86,67

Table 1

The table shows the results obtained by the two linear discriminant techniques on real sequences. The rows show the classification percentage obtained on the 7 different sequence concerning different persons. The columns indicated as "IDX" is related to those classifications that returned the identity of one the 40 people extracted from the Feret thus not belonging to the watch-list.

classifiers the performances increase. In particular, in this last case the correct classification rate is around the 70%.

To assess the performance of the face classifier in a really tough situation, a final test has been executed on a heavy occluded face patterns. In particular, sequences of the AVIRES member wearing a full face helmet have been acquired for such purposes. In this case, just the results of the RFLD classifier on the eyes patterns have been considered. On such sequences, the values reported in Table 3 show a correct classification rate that is slightly above the 50%. In this situation, the appearance of the detected patterns resulted to be really different to the training patterns. For this reason, the system's performance results so limited.

	Whole Face							Eyes Region						
x \ c	ID1	ID2	ID3	ID4	ID5	ID6	IDX	ID1	ID2	ID3	ID4	ID5	ID6	IDX
ID1	58,43	14,61	11,24	3,37	2,25	0,00	10,11	69,66	7,87	8,99	3,37	2,25	0,00	7,87
ID2	12,50	39,84	17,97	3,91	3,13	2,34	20,31	8,59	68,75	9,38	2,34	2,34	0,00	8,59
ID3	12,81	8,54	41,99	4,27	11,03	4,98	16,37	5,69	3,91	70,82	0,71	1,07	1,42	16,37
ID4	8,94	5,69	4,88	29,27	13,01	15,45	22,76	0,81	0,00	0,00	69,11	7,32	8,13	14,63
ID5	0,00	0,00	0,00	3,62	35,75	27,15	33,48	0,00	0,45	0,45	7,69	71,49	8,60	11,31
ID6	0,89	0,00	2,68	9,82	18,75	41,96	25,89	0,89	0,00	2,68	8,93	6,25	64,29	16,96

Table 2

The results reported in the table show the performance of the RFLD based face classifier on two different types of patterns. On the left side the results obtained by using the whole-face patterns occluded by the scarf are shown. Right side presents the result obtained by using a small area surrounding the eyes.

	Helmet- Eyes Region						
x \ c	ID1	ID2	ID3	ID4	ID5	ID6	IDX
ID1	57,41	8,33	4,63	4,63	0,93	0,93	23,15
ID2	8,87	58,06	12,10	0,81	0,81	0,00	19,35
ID3	9,70	2,42	56,36	3,64	2,42	9,70	15,76
ID4	1,63	0,00	1,09	54,89	10,87	9,24	22,28
ID5	0,00	0,00	0,00	1,00	59,70	22,39	16,92
ID6	0,84	0,00	2,52	3,36	15,13	59,66	18,49

Table 3

The table presents the results obtained by the RFLD technique on patterns obtained by cropping a small area surrounding the eyes of subjects wearing an helmet.

5.2 Speaker Recognition

To confirm the effectiveness of de-noising methods described in the previous section, we performed two speaker recognition tasks on:

Corpus	Condition	Noise	Den-oise	Error rate
ELSDSR	a1	No	No	4.50%
ELSDSR	a2	Yes	No	18.20%
ELSDSR	a3	Yes	Yes	11.40%
AVIRES	b1	Yes	No	22.70%
AVIRES	b2	Yes	Yes	13.60%

Table 4

Experimental results under five different conditions.

a. ELSDSR Corpus. To all clean signals $x_C(t)$ 10 dB white-noise $x_N(t)$ was added, in order to synthesize the stimuli $x(t)$.
b. AVIRES Corpus.

We conducted experiments under the following five conditions: (a1) recognition of ELSDSR $x_C(t)$ data; (a2) recognition of ELSDSR $x(t)$ data; (a3) recognition of ELSDSR $x(t)$ data using speech de-noising (see Sec. 3.1); (b1) recognition of AVIRES data; (b2) recognition of AVIRES data using speech de-noising (see Sec. 3.1).

Table 4 lists experimental results. The comparison of conditions (a1) and (a2) shows that the error rate increased significantly from 4.5% to 18.2%. This increase indicates an evident influence of the noise. When speech de-noise is used, the error rate decreased from 18.2% to 11.4%, which correspond to a reduction in error rate of about 6.8%. Using the AVIRES corpus (in which the data are affected by noise and the speakers using scarf and full face helmet) the reduction in error rate is of about 9.10%.

In order to evaluate the performance of our system, we adopted the DET (Detection Error Trade-off) curves. They are a variant of the ROC curves

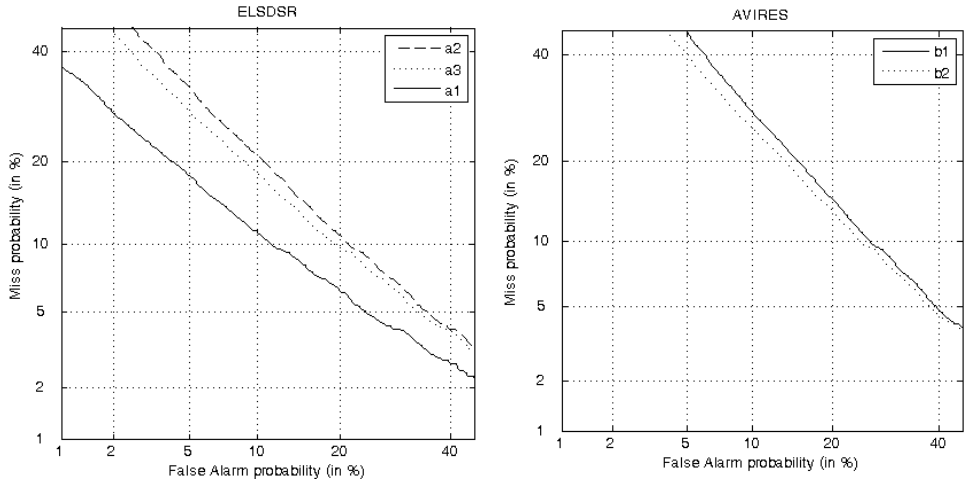

Figure 6. *The charts plot the miss probability Vs. false alarm probability. (a) presents such an evaluation on a standard dataset (i.e. ELSDSR). Precisely the (a1), (a2) and (a3) conditions are presented respectively by the continuous, dashed and dotted lines. (b) presents the same evaluation for the AVIRES corpus in which (b1) and (b2) are respectively plotted with continuous and dotted.*

such that the error rates are plotted on both axes, giving uniform weight to both types of error. This allows to better distinguish the different performances even when they are similar [46]. Fig. 6(a) and 6(b) respectively show the DET curves of the experiments {(a1), (a2), (a3)} and {(b1), (b2)}. It is worth noticing how, in both cases, the system performance increases when the denoise algorithm is used (dotted curve).

In order to compare the results of the speaker recognition task with the results of video recognition task, we calculated the probabilities that a speech Y is from the speaker S (for each speaker and each speech) in the experiment under condition b2 (see Tab. 5).

		ID1	ID2	ID3	ID4	ID5	ID6	IDX
Normal	ID1	81,30%	5,90%	7,15%	0,90%	1,90%	2,10%	0,75%
	ID2	5,20%	68,00%	8,80%	0,85%	2,05%	2,00%	13,10%
	ID3	5,23%	8,77%	80,00%	1,00%	1,90%	1,60%	1,50%
	ID4	0,80%	0,50%	0,90%	77,90%	3,14%	3,12%	13,64%
	ID5	7,34%	2,10%	1,10%	2,50%	48,00%	5,20%	33,76%
	ID6	1,54%	1,22%	0,80%	1,00%	17,00%	65,00%	13,44%
Scarf	ID1	55,00%	13,87%	16,90%	1,13%	5,87%	4,90%	2,33%
	ID2	9,00%	68,00%	9,60%	1,80%	5,50%	2,40%	3,70%
	ID3	6,67%	8,54%	76,00%	0,90%	2,40%	2,10%	3,39%
	ID4	0,70%	0,67%	2,20%	60,10%	9,80%	12,10%	14,43%
	ID5	1,60%	2,90%	1,40%	3,30%	44,00%	9,77%	37,03%
	ID6	2,20%	2,34%	1,30%	1,00%	18,60%	53,00%	21,56%
Helmet	ID1	61,00%	12,80%	14,85%	1,11%	3,88%	2,10%	4,26%
	ID2	5,40%	43,66%	47,82%	0,22%	0,45%	1,00%	1,45%
	ID3	7,04%	7,65%	77,80%	0,87%	1,95%	1,50%	3,19%
	ID4	0,77%	1,10%	1,22%	50,56%	9,33%	11,30%	25,72%
	ID5	3,45%	5,70%	1,22%	7,60%	22,00%	15,22%	44,81%
	ID6	1,10%	0,70%	0,90%	2,20%	18,30%	37,75%	39,05%

Table 5

Probabilities that a speech Y is from the speaker S (for each speaker and each speech) in the experiment under condition b2.

5.3 Audio-Video Recognition

To test the integration of the video and speaker classifiers, for each subject of the AVIRES corpus we have extracted sequences of $22s$ each belonging to the

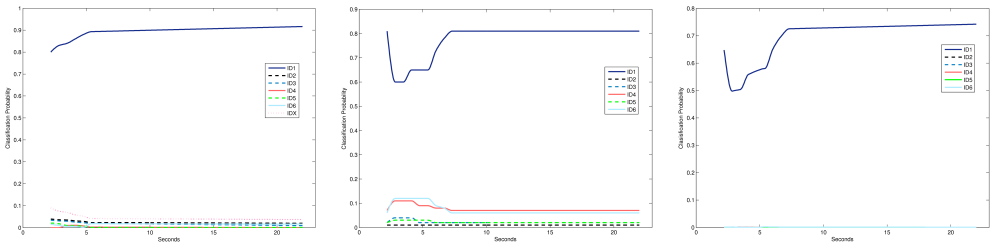

Figure 7. *Classification results obtained on a normal sequences for the* $ID1$ *subject. (a) Plots the video classification probability scores over time computed by using the temporal integration. 7(b) Plots the speaker classification probability scores over time computed every* 1100*s. 7(c) Plots the classification probability scores obtained by the integration rule.*

three different situations: normal, scarf, helmet.

In Figure 7 results obtained for the $ID1$ normal sequences are presented. Precisely, 7(a) presents the probability scores associated to any possible class on pattern belonging to class $ID1$. The temporal integration for video scores is adopted. It is interesting to notice how, even if the classifier has 86% of correct detection in this situation, the temporal integration filters out single images classification errors by keeping the correct classification probability around 90%. In addition, it can be noted the the probabilities associated to the other classes (i.e. the error) fall all below 5%. Figure 7(b), shows respective values computed by the speaker classifier. Due to development limitations of the de-noising algorithm the audio track has been subdiveded in segments of $2200ms$. To relax such a constraint we have adopted a sliding window that extracts $2200ms$ long samples every $1100ms$. This allowed to have twice the number of speaker classifications scores (i.e. every $1100ms$ against $2200ms$).

The results obtained by the speaker classifier in the normal condition can be considered good even though the algorithm showed a problem in the classification of the second pattern. Such a problem yields to a reduction of the

correct classification probability in correspondence of the second score. In addition, due to the Kalman filtering such an error is propagated for the following 5 classifications. Anyway, even the speaker recognition in this condition performs very good. As matter of fact, the correct classification probability is maintained around the 80%, falling to a fair 60% only in correspondence of the second pattern. The classification probabilities computed for the wrong classes are kept below 10%. Finally, if we look at the integration results presented in Figure 7(c), it can be noted how the error introduced by the speaker classifier is still propagated but keeping the final classification probability around the 70% yet. In addition, the integration of the audio and video scores allowed to filter out the wrong clasisfication probabilities. As matter of fact, while the ratio between correct association and wrong associations are 17 and 9 respectively for the video and audio classifiers, the integration shows a ratio of 75. Hence, the discrimination between correct and wrong classification improved of a factor 4 and 8 with respect to video and audio discrimination respectively. In Figure 8 an overall evaluation for all the subjects belonging to the AVIRES corpus is presented. It can be noted how in this situation the video classification would be sufficient since no occlusions occur. Anyhow, by integrating the two classifiers results even if the correct classification probability decreases for all the subjects, the correct versus wrong classifications ratio increases. Such an increase is around 6% on the average.

As done for the normal situation for which non occlusions occur, we have tested the algorithm for the scarf and helmet conditions. Figure 9 and 10 show the obtained correct classification probability scores respectively for the scarf and helmet situations. In these two conditions, while the video scores (9(a) and 10(a)) are not as reliable as those computed for the normal situation, the

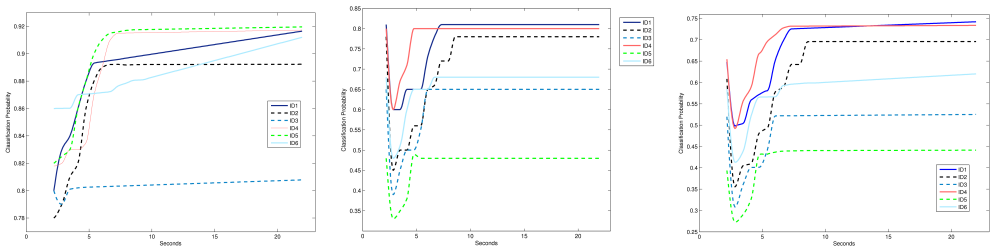

Figure 8. *No occlusions - Correct classification results obtained by the video (a), speaker (b) and computed by the integration rule (c).*

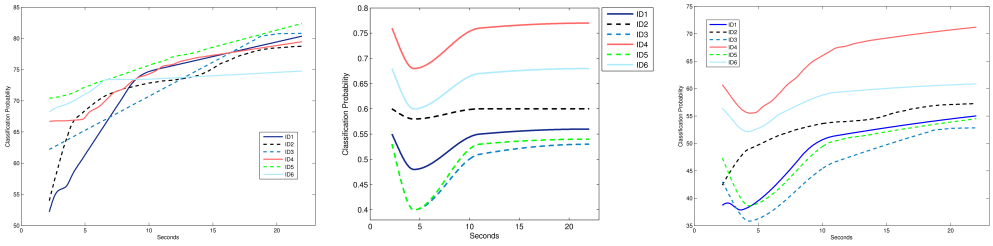

Figure 9. *Scarf - Correct classification results obtained by the video (a), speaker (b) and computed by the integration rule (c).*

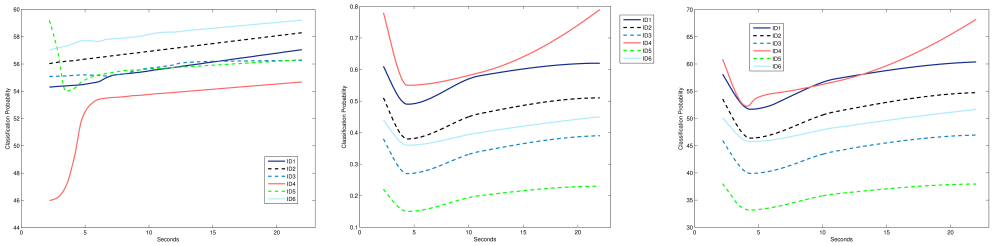

Figure 10. *Helmet Correct classification results obtained by the video (a), speaker (b) and computed by the integration rule (c).*

speaker scores (9(b) and 10(b)) obtains similar performances. This allow the system to keep almost the same correct classification probabilities when video and audio classification is integrated (9(b) and 10(c)).

Such performances, on the three situations (no-occlusions, scarf, helmet), generate the classification error chart plotted in Figure 11. In such a chart, the mean classification errors computed on the integration results are considered. It is worth noticing how the overall error is kept below the 20% in all the three

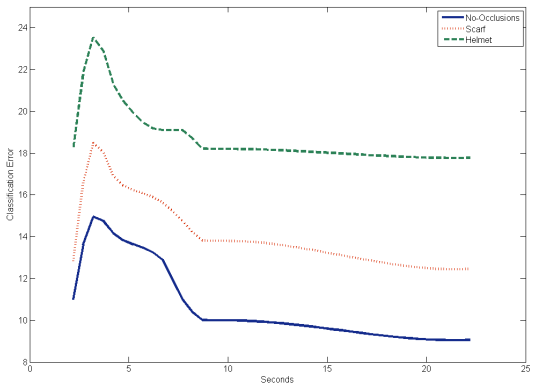

Figure 11. *The mean classification error computed on all the sequences belonging to the same conditions are plotted.*

situations. This shows how the fusion of video and audio scores represents a reliable way to grant access for restricted zones.

Conclusions

In this paper we presented different techniques for the automatic recognition of non-cooperative persons that try to get access to a restricted zone. In particular, such an objective has been achieved by developing video and audio based classifiers. In the context of visual recognition, we have proposed a technique able to detect the face or parts of it even in a non frontal pose or when the person is wearing clothes that occlude some features. In order to classify such patterns we have selected two different linear discriminant techniques like R-LDA and RFLD to project the patterns in a new space. On this space a Mahalonobis distance is computed to provide a frame-by-frame classification probability on the new space. The obtained probabilities are therefore temporally associated by using an independent probability assumption to obtain an improvement of the robustness . In the context of speaker recognition,

we described a noise reduction system based on the Extended Kalman Filter (EKF) optimized for speech de-noising by means of some improvements about the bootstrap procedure, the stability check and the Variable Forgetting Factor. We used this method, in noisy environments, in combination with a conventional speaker recognition system, based on Gaussian mixture models and Mel Frequency Cepstral Coefficients (MFCCs) as features.

Experiments have been proposed by using two different corpora. Video and speaker recognition applied to such corpora show some limitations due to the non cooperative context. Anyway, by adopting the temporal video classification and integrating it with the speaker classification, the system is able to obtain interesting results where the drawbacks of each of the two methods are smoothed by the qualities of the other.

References

[1] A. Jain, Biometric recognition: how do i know who you are?, in: Signal Processing and Communications Applications Conference, 2004, pp. 3–5.

[2] E. Osuna, R. Freund, F. Girosi, Training support vector machines: an application to face detection, in: Conference on Computer Vision and Pattern Recognition, Washington D.C., 1997, pp. 130–136.

[3] H. Rowley, S. Baluja, T. Kanade, Neural network-based face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1) (1998) 23–38.

[4] F. Samaria, S. Young, Hmm-based architecture for face identification, Image and Vision Computing 12 (8) (1994) 537–543.

[5] P. Viola, M. Jones, Robust real-time face detection, International Journal of

Computer Vision 57 (2) (2004) 137–154.

[6] M. Yang, D. Kriegman, N. Ahuja, Detecting faces in images: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (1) (2002) 34–58.

[7] W. Zhao, R. Chelappa, P. Phillips, A. Rosenfled, Face recognition: A literature survey, ACM Computer Survey 35 (4) (2003) 399–458.

[8] P. Phillips, H. Moon, S. Rizvi, P. Rauss, The feret evaluation methodology for face recognition algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 1090–1104.

[9] C. Xiang, X. A. Fan, T. H. Lee, Face recognition using recursive fisher linear discriminant, IEEE Transactions on Image Processing 15 (8) (2006) 2097–2105.

[10] A. Hadid, M. Pietikäinen, From still image to video-based face recognition: an experimental analysis, in: Sixth IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, 2004, pp. 813–818.

[11] S. Zhou, V. Krueger, R. Chellappa, Probabilistic recognition of human faces from video, Computer Vision and Image Understanding 91 (2003) 214–245.

[12] A. Abate, M. Nappi, D. Riccio, G. Sabatino, 2d and 3d face recognition: A survey, Pattern Recognition Letters 28 (2007) 1885–1906.

[13] C. Micheloni, G. Foresti, Image acquisition enhancement for active video surveillance, in: IAPR International Conference on Pattern Recognition, Cambridge, UK, 2004, pp. 271–275.

[14] I. Visentini, C. M. ang G.L. Foresti, Tuning asymboost cascades improves face detection, in: IEEE International Conference on Image Processing, Vol. 4, San Antonio, Texas U.S.A., 2007, pp. 477–480.

[15] Z. Zhang, M. Li, S. Z. Li, H. Zhang, Multi-view face detection with floatboost, in: WACV '02: Proceedings of the Sixth IEEE Workshop on Applications of

Computer Vision, IEEE Computer Society, Washington, DC, USA, 2002, p. 184.

[16] J. Lu, K. Plataniotis, A. Venetsanopoulos, Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition, Pattern Recognition Letters 26 (2) (2005) 181–191.

[17] D. Reynolds, Speaker identification and verification using gaussian mixture speaker models, in: ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, 1994, pp. 27–30.

[18] D. Reynolds, T. Quatieri, R. Dunn, Speaker verification using adapted gaussian mixture models, Digital Signal Processing 10 (2000) 19–41.

[19] J. Lim, A. Oppenheim, All-pole modeling of degraded speech, IEEE Trans. Acoust., Speech, and Signal Process 26 (3) (June 1978) 197–210.

[20] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE Trans. Acoust., Speech, and Signal Process 32 (6) (1984) 1109–1121.

[21] D. Tsoukalas, J. Mourjopoulos, G. Kokkinakis, Speech enhancement based on audible noise suppression, IEEE Trans. Acoust., Speech, and Signal Process 5 (6) (Nov. 1997) 497–514.

[22] N. Virag, Single channel speech enhancement based on masking properties of the human auditory system, IEEE Trans. Acoust., Speech, and Signal Process 7 (2) (Mar. 1999) 126–137.

[23] M. Niedźwiecki, K. Cisowski, Adaptive scheme for elimination of broadband noise and impulsive disturbances from AR and ARMA signals, IEEE Trans. Signal Process 44 (3) (1996) 967–982.

[24] N. Ma, M. Bouchard, R. A. Goubran, Speech enhancement using a masking

threshold constrained Kalman filter and its heuristic implementations, IEEE Trans. Speech, Audio, and Language Process 14 (1) (Jan 2006) 19–32.

[25] V. Grancharov, J. Samuelsson, B. Kleijn, On casual algorithms for speech enhancement, Trans. Audio, Speech, and Language Process. 14 (3) (May 2006) 273–276.

[26] P. A. A. Esquef, V. Valimaki, M. Karjalainen, Restoration and enhancement of solo guitar recordings based on sound source modeling, J. Audio Eng. Soc. 50 (4) (2002) 227–236.

[27] M. Niedźwiecki, Identification of time-varying processes in the presence of measurement noise and outliers, in: Proc. 11th IFAC Symposium on System Identification, 1997, pp. 1765–1768.

[28] B. D. O. Anderson, J. B. Moore, Optimal filtering, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[29] J. D. Markel, A. K. Gray, Linear prediction of speech, Springer Verlag, Berlin, 1975.

[30] M. Niedźwiecki, Identification of Time-Varying Processes, Wiley, New York, 2000.

[31] A. Bari, S. Canazza, G. D. Poli, G. Mian, Toward a methodology for the restoration of electro-acoustic music, J. New Music Research 30 (4) (2001) 365–374.

[32] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust., Speech, and Signal Process 28 (1980) 357–366.

[33] C. Broun, W. Campbell, D. Pearce, H. Kelleher, Speaker recognition and the etsi standard distributed speech recognition front-end, in: Speaker Odyssey - The Speaker Recognition Workshop, Crete, Greece, 2001, pp. 121–124.

[34] H. Melin, Automatic speaker verification on site and by telephone: methods, applications and assessment, Ph.D. thesis, KTH, Stockholm (December 2006).

[35] A. Rosenberg, F. Soong, Evaluation of vector quantization talker recognition system in text independent and text dependent modes, Computer Speech and Language 11 (3) (1987) 873–876.

[36] D. Burton, Text-dependent speaker verification using vector quantization source coding, IEEE Trans. Acoust., Speech, and Signal Process 35 (2) (1987) 133–143.

[37] J. Naik, L. Netsch, G. Doddington, Speaker verification over long distance telephone lines, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Glasgow, UK, 1989, pp. 524–527.

[38] M. Savic, S. Gupta, Variable parameter speaker verification system based on hidden markov modeling, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Albuquerque, NM, USA, 1990, pp. 281–284.

[39] D. Reynolds, A gaussian mixture modeling approach to text-indipendent speaker identification, Ph.D. thesis, Georgia Institute of Technology, Atlanta, Ga, USA (September 1992).

[40] D. Reynolds, Speaker identification and verification using gaussian mixture speaker models, Speech Communication 17 (1) (1995) 91–108.

[41] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, Score normalization for text-independent speaker verification systems, Digital Signal Processing 10 (2000) 42–54.

[42] M. Niedźwiecki, Steady-state and parameter tracking proprieties of self-tuning minimun variance regulators, Automatica (1989) 597–602.

[43] L. Feng, L. K. Hansen, A new database for speaker recognition, Tech. rep., Informatics and Mathematical Modelling, Technical University of Denmark,

Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby (2005).
URL http://www2.imm.dtu.dk/pubdb/p.php?3662

[44] S. Linville, Vocal Aging, Singular Thomson Learning, San Diego, 2001.

[45] I. Murray, L. Arnott, Toward the simulation of emotion in synthestic speech: A review of the literature on human vocal emotion, Journal of The Acoustical Society of America 93 (2) (1993) 1097–1108.

[46] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The det curve in assessment of detection task performance, in: Proceedings of EuroSpeech 1997, Vol. 4, 1997, pp. 1895–1898.