ENTERFACE'08, AUGUST 4TH - AUGUST 29TH, ORSAY, FRANCE. FINAL PROJECT REPORT

Activity-related Biometric Authentication

G. Ananthakrishnan, H. Dibeklioğlu, M. Lojka, A. Lopez, S. Perdikis, U. Saeed, A.A. Salah, D. Tzovaras, A. Vogiannou

***Abstract*—This project aims at developing a biometric authentication system exploiting new features extracted by analysing the dynamic nature of various modalities, including motion analysis during ordinary tasks performed in front of a computer, analysis of speech, continous face and facial movement analysis, and even patterns for grasping objects. We test the potential and contribution of each of these modalities for biometric authentication in the face of natural, uncontrolled environments, as well as their fusion.**

***Index Terms*—Biometric authentication, activity recognition, face recognition, motion analysis, speaker recognition, audio based event recognition**

I. Introduction

THIS project attempts to address the limitations of unimodal biometrics by deploying activity-related multimodal biometric systems that integrate the evidence presented by multiple sources of information. Therefore, the combination of a number of independent modalities is explored to overcome the possible restrictions set by each modality. With a simple sensor setup, we aim at more robust biometric identification through the fusion of physiological, behavioral and soft biometric modalities, keeping also in mind the unobtrusiveness and comfort of the subject.

The term behavioral biometrics refers to Person Recognition using shape based activity signals (gestures, gait, full body and limb motion) or face dynamics. Activity-specific signals [6], [23] provide the potential of continuous authentication, but state-of-the-art solutions show inferior performance compared to static biometrics (fingerprints, iris). This drawback could hopefully be eliminated by the inferential integration of different modalities.

Behavioral information from face videos for person recognition may also be investigated in order to exploit the underlying temporal information in comparison to image-based recognition [39]. Methods for person recognition from face dynamics can be classified into holistic methods (head displacements and pose evolution [30]), feature-based methods (exploitation of individual facial features [12]) and hybrid methods [14]. Various probabilistic frameworks have been proposed in recent works, usually employing a Bayesian Network (Hidden Markov Models, Coupled and Adaptive HMMs, etc.) as the mathematical model for recognition [35].

Soft biometrics (gender, height, age, weight etc.) are believed to be able to significantly improve the performance of a biometric system in conjunction with conventional static biometrics [22], yet their exploitation remains an open issue. Microphones for voice recognition, sound based sensors for monitoring activities or other modalities could also be considered.

This report, as well as the source code for the software developed during the project, is available online from the eNTERFACE'08 web site: www.enterface08.limsi.fr.

G. Ananthakrishnan is with Royal Institute of Technology, SWEDEN. E-mail: agopal@kth.se.

H. Dibeklioğlu is with Perceptual Intelligence Laboratory, Department of Computer Engineering, Boğaziçi University, 34342 Bebek, Istanbul, TURKEY. E-mail: hamdi.dibeklioglu@cmpe.boun.edu.tr.

M. Lojka is with Technická Univerzita v Košiciach, SLOVAKIA. E-mail: martin.lojka@tuke.sk.

A. Lopez is with Technical University of Catalonia, Barcelona, SPAIN. E-mail: alopez@gps.tsc.upc.edu.

S. Perdikis, A. Vogiannou and D. Tzovaras are with CERTH/ITI, GREECE. E-mail: {perdik,tvog }@iti.gr.

U. Saeed is with EURECOM, FRANCE. E-mail: usman.saeed@eurecom.fr.

A.A. Salah is with Centrum Wiskunde & Informatica, 1090 GB Amsterdam, THE NETHERLANDS. E-mail: a.a.salah@cwi.nl.

In this report, we look at some of these modalities in a specific fixed-seat pilot. Our experimental setup is described in Section II, including the details of the collected database. The individual modalities are investigated in separate sections, starting with model-based motion analysis in Section III, which tracks the user via calibrated cameras during ordinary activities. The sounds that ensue during these activities are analysed for robust activity classification. This part is exposed in Section IV. Once speech is detected among the sound events, it can be further used for authentication. Section V deals with speaker authentication. Our model flexibly integrates data coming from seemingly unrelated modalities. Section VI exemplifies this by making use of an advanced interface for recognizing activity, namely a Cyberglove, which is used to collect and analyse grasping patterns. The more common face modality is used to serve as a benchmark. Continuous authentication from captured static face images is explained in Section VII, and the optical-flow based analysis of facial motion for authentication is detailed in Section VIII. Section IX builds on the motion analysis to recognize types of activities, and evaluates the authentication potential of each of these activities.

The mathematical framework we establish here is employed to seamlessly integrate an arbitrary number of sources that provide partial authentication information. Our experimental results are given in Section XI. The report concludes with a discussion of these results and on possible future directions in Section XII.

II. The Experimental Setup

The proposed biometric system is evaluated in a fixed - seat office pilot, where the user is able to move his arms, head and torso and manipulate objects on a desk while seated. This experimental setup selection serves multiple aspects of the problem of activity-related biometric authentication:

- **It is portable and easy to setup**
- **It can be part of a normal authentication system scenario (e.g. secured indoor premises)**
- **It can easily incorporate all the equipment for selected modalities**
- **An office environment is involved in many work - related activities, which makes the pilot ideal for testing the activity - related authentication module**
- **It is fully unobtrusive to the user**

The selected pilot consists of a desk upon which a number of objects is placed, in stable predefined positions. This constraint implies a static environment, which slightly affects the generality of the setup, but significantly facilitates the activity recognition task. The objects are: a) desk phone, b) glass (on a pad), c) keyboard, d) mouse, e) computer screen, f) pencil (in a pencil case) g) a piece of paper for writing. The sensorial equipment is as inexpensive and unobtrusive as possible. It comprises of three Logitech QuickCam webcams (two for body motion tracking and one for continuous face authentication and facial motion analysis) and a regular low - budget microphone. Two cameras are mounted on the desktop screen facing the user (these are the frontal motion tracking camera and the face camera, which is zoomed on the user's head area), while the third camera (lateral motion tracking camera) is placed on a tripod on the left side of the desk. The microphone is mounted on the desk, next to the keyboard. Fig. 1 illustrates the actual pilot setup.

A. Recording Scenario and Data Gathering

Within the project a database of 15 persons performing a number of actions has been recorded. Each person was asked to execute six actions in a particular order, responding to the environmental stimuli (phone ringing, instructions on the screen or on a writing form). A recording scenario has been prepared so as to enhance the database's consistency, to meet the requirements and constraints of every modality and to ensure the user's concentration and relaxation, so that he performs the required

Fig. 1. The pilot setup, shown during one of the recordings. The frontal cameras are mounted on the display, and the side camera is mounted on a pod to the left of the subject.

actions in his natural (and therefore consistent) way. The six recorded actions were:

- **Mouse manipulation (playing a computer game)**
- **Phone Conversation (real dialogue with a team member)**
- **Typing in the keyboard (filling in a given questionnaire)**
- **Writing (filling in a questionnaire in a writing form)**
- **Drinking (taking the cup, and leaving it back to its place)**
- **Reading (specific texts provided in the screen).**

Every session consisted of one repetition of the six actions and 10 **sessions were recorded for each user in order to provide enough training and testing data for all the modalities. The database size was limited to** 15 **persons due to limited time available for recordings.**

During the data gathering users were asked to act in their natural way, without any further instructions or constraints. The selected activities are common work - related activities involving usual office objects, there was no previous knowledge about their suitability for authentication. The evaluation of their discriminative power is among the objectives of this project.

III. Model-Based Motion Analysis

Markerless human motion capture is a challenging problem that involves the estimation of a high-dimensional configuration of a three-dimensional non-rigid and self-occluding object. Since a wide range of applications are derived from the unobtrusive characterization of human activity, this research area has recently undergone several advances due to the yielded interest.

A common approach is to consider an articulated body model with several degrees of freedom per joint, depending on the complexity of the possible poses and the quality of the available data. This representation implies the use of kinematic constraints on the motion. Additional assumptions and motion constraints can be adopted at the cost of generality of the solution which we intend to preserve. To this end, Particle Filters [2] have become a relevant technique due to their ability to handle multi-modal non-linear and non-Gaussian distributions. Several approaches such as partitioned sampling [37], hierarchical sampling [41] and annealing particle filter [15] have been developed to cope with high-dimensional limitations of the classical Condensation algorithm [21].

We present a particular implementation of the annealing particle filter for a simplified body model in order to retrieve the human body poses of a subject performing different actions in a multi-view scenario. We propose simplifications of the body tracking problem without almost no loss of generality in the given pilot and with the capability of coping with realistic scenarios.

TABLE I
Articulated Body Model Joints

Angle	Joint	Rotation Axis	Range
θ_1	Base of the Neck	**y**	$[-\frac{\pi}{4}, \frac{\pi}{4}]$
θ_2	Right Shoulder	**x**	$[-\frac{\pi}{4}, \pi]$
θ_3	Left Shoulder	**x**	$[-\frac{\pi}{4}, \pi]$
θ_4	Right Shoulder	**y**	$[-\frac{\pi}{4}, \pi]$
θ_5	Left Shoulder	**y**	$[-\frac{\pi}{4}, \pi]$
θ_6	Right Shoulder	**z**	$[-\frac{\pi}{4}, \frac{\pi}{2}]$
θ_7	Left Shoulder	**z**	$[-\frac{\pi}{4}, \frac{\pi}{2}]$
θ_8	Right Elbow	**y**	$[0, \pi]$
θ_9	Left Elbow	**y**	$[0, \pi]$

A. Body Model

A simplistic articulated body model will fulfill the requirements of the scenario presented in section II. This model is based on the kinematic chain framework and comprises a set of joints. In our case, this set of joints are the base of the neck, shoulders and elbows. Every joint has a maximum of three degrees of freedom according to the complexity of the motions that we want to capture. Each degree of freedom is represented by an axis of rotation defined in a default body configuration, where all the angles are set to zero (see fig. 2). The range of joint angles is also defined according to this default body pose. In our model, a total of nine degrees of freedom are defined (see table I). In order to set the model in a world position, a three-dimensional coordinate system built with the base of the neck as origin and a body orientation are defined. Our model reference point is set to be the base of the neck. Therefore, our body model defines a thirteen-dimensional state vector:

$$\mathbf{x_t} = \{x_0, y_0, z_0, \theta_0, ..., \theta_9\} \tag{1}$$

Angle θ_0 is the orientation of the whole body model while all the other angles are designed following basic kinematic constraints. The use of angles ensures a compact representation in front of a state defined by only 3D coordinates. Knowing the limbs' dimensions we can go from a set of angles to Cartesian coordinates by means of exponential twists formulation [7]; every point of interest can be computed from its initial location with respect to the reference point in the default body configuration and the product of the exponential maps affecting the motion of this point:

$$p(\mathbf{x_t}) = \prod_i \mathbf{M}_i(\mathbf{x_t}) p_0 \tag{2}$$

$$\mathbf{M}_i = \begin{bmatrix} \mathbf{R_i}(\mathbf{x_t}) & \mathbf{t_i}(\mathbf{x_t}) \\ 0 & 1 \end{bmatrix} \tag{3}$$

where $p(\mathbf{x_t})$ represents a point of interest as a function of the state vector, that encodes model position, model orientation and joint angles, and $\mathbf{M}_i(\mathbf{x}_t)$ is the exponential map in the chain where p is found. The exponential map comprises the rotation matrix R and the translation vector t. The whole notation is being presented in homogeneous coordinates due to its compactness.

B. Particle Filter

Particle Filters (PF) [2] are recursive Bayesian estimators derived from Monte Carlo sampling techniques which can handle non-linear and non-Gaussian processes. Commonly used in tracking problems, they are used to estimate the posterior density $p(\mathbf{x}_t|\mathbf{z}_t)$ by means of a set of

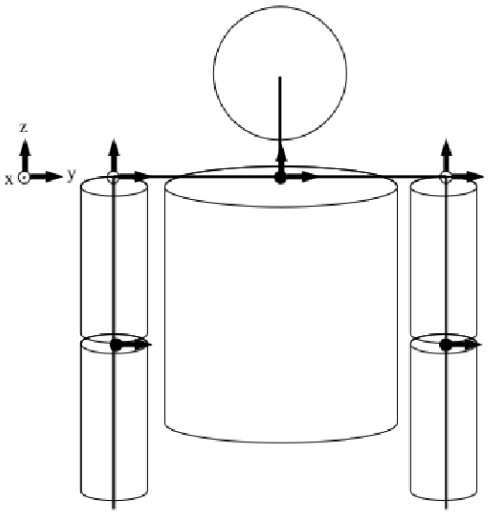


Fig. 2. Simple articulated model for body tracking

N_s **weighted samples or particles. Given a Bayesian recursive estimation problem:**

$$p(\mathbf{x}_{0:t}|\mathbf{z}_{1:t}) = \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})} p(\mathbf{x}_{0:t-1}|\mathbf{z}_{1:t-1}) \quad (4)$$

we want to draw samples from the posterior such that:

$$p(\mathbf{x}_{0:t}|\mathbf{z}_{1:t}) \approx \sum_{i}^{N_s} w_t^i \delta(\mathbf{x}_t - \mathbf{x}_t^i) \quad (5)$$

where w_t^i is the weight associated to the i-th particle. This discrete approximation of the posterior requires the weights evaluation. This is done by means of the importance sampling principle [16], with a probability density function (pdf) $q(\mathbf{x}_{0:t}|\mathbf{z}_{1:t})$ from which we can generate samples that can be evaluated with the posterior (up to proportionality). Applying the importance sampling principle in Eq. 4:

$$w_t^i \propto \frac{p(\mathbf{x}_{0:t}|\mathbf{z}_{1:t})}{q(\mathbf{x}_{0:t}|\mathbf{z}_{1:t})} \quad (6)$$

$$w_t^i \propto \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})q(\mathbf{x}_{0:t}|\mathbf{z}_{1:t})} p(\mathbf{x}_{0:t-1}|\mathbf{z}_{1:t-1}) \quad (7)$$

and choosing this importance distribution in a way that factors appropiately we have:

$$w_t^i \propto \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{0:t-1}|\mathbf{z}_{1:t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})q(\mathbf{x}_t|\mathbf{x}_{0:t-1},\mathbf{z}_t)q(\mathbf{x}_{0:t-1}|\mathbf{z}_{1:t-1})} \quad (8)$$

$$w_t^i \propto w_{t-1}^i \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})} \quad (9)$$

Moreover, if we apply the Markov assumption the expression is simplified regarding the fact that observations and current state only depend on the previous time instant. Therefore, the PF is a sequential propagation of the importance weights.

Two major problems affect the PF design. The first is the choice of the importance distribution. This is crucial since the samples drawn from $q()$ must hit the posterior's typical set in order to produce a good set of importance weights. It has been shown in [16] that $q(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{z}_t) = p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{z}_t)$ is optimal in terms of variance of the weights. The second problem deals with particle degeneracy in terms of variance of the weights. After several iterations the majority of the particles have negligible weights and as a consequence of this the estimation efficiency decays. An effective measure for the particle degeneracy is the survival rate [34] given by:

$$\alpha = \frac{1}{N_s \sum_{i=1}^{N_s} (w_t^i)^2} \quad (10)$$

In order to avoid the estimator degradation the particle set is resampled. After likelihood evaluation a new particle set must be drawn from the posterior estimation, hence particles with higher weights are reproduced with higher probability. Once the new set has been drawn all the weights are set to $\frac{1}{N_s}$, leading to a uniformly weighted sample set concentrated around the higher probability zones of the estimated posterior.

The Sampling Importance Resampling (SIR) Particle Filter proposed by Gordon et. al [18] is a method commonly used in computer vision problems. It's characterized by applying resampling at every iteration and by defining the importance distribution as the prior density $p(x_t|x_{t-1})$. By substituting this importance density in 8, it's easy to realize that weight computation only depends on the likelihood. Consequently, the design of the particle filter is basically a problem of finding an appropiate likelihood function.

C. Likelihood Evaluation

In computer vision problems probability density functions usually are not directly accessible, thus an observation model is required to approximate the likelihood function. It is necessary to determine which image features are more correlated with the true body configuration. Therefore, finding the appropiate likelihood approximation involves both image and body model. Deutscher et al. [15] proposed a matching of the model projection with foreground segmentation and edges. Their flesh model consists of conic sections with elliptical cross-sections surronding virtual skeleton segments. Raskin et al. [50] add the body part histogram as an additional feature. Other authors use Visual Hull approaches [27] to work with voxel data. In that case, they can use three-dimensional flesh models, like ellipsoids [40] or three-dimensional Gaussian mixtures [9].

Our challenge is to produce a likelihood approximation able to deal with moving objects, clothing, limited number of views and low frame rate. In our approach we should not rely on a 3D reconstruction because only a few views are available, thus a projection of the model onto the images is required. Our proposal is to avoid the computational cost of projecting the whole set of sampling points of a 3D flesh model by projecting a reduced set of points per body part. Our flesh model will be set of cylinders around all the skeleton segments except the head, which will be modelled by a sphere (see Fig. 2). Therefore, our reduced set of projected points will be defined by the vertices of the trapezoidal section resulting from the intersection of a plane, approximately parallel to the image plane, with the cylindric shapes modelling the limb (or spherical shape in the case of the head).

To define an intersecting plane for a given cylinder, we compute the vectors going from the camera center towards each one of the limit points of the limb. Then the cross product of these vectors with the one defined by the limb itself is computed to determine two normal vectors that lie on the intersecting plane and along which we will find the key points to project. The head template is handled with a similar procedure using as limb vector the one going from the body model reference point to the head center. The norm of the cross product, as well as the area of the projected trapezoid, can be used as a quality measure in order to determine whether the limb is properly aligned with the view (this does not apply for the head). If this quality measure is above a certain threshold, we can change the trapezoidal projected shape by a circle or an ellipse. However, in our scenario the views are set so that they capture good limb alignments in most of the frames, thus we can obviate the computation of this measure.

Regarding the image features, we have seen that common likelihood approximations like [15] do not perform well in our scenario with the described body model. We propose modifications on this approximation while keeping common features that are easy to extract, like foreground silhouettes, contours and detected skin. We extract foreground silhouettes by means of a background learning technique based on Stauffer and Grimson's method [56]. A single multivariate Gaussian $\mathcal{N}(\mu_t, \mathbf{\Sigma}_t)$ with diagonal covariance in the RGB space is used to model every pixel value $\mathbf{I_t}$. The algorithm learns the background model for every pixel using a set of background images and then, for the rest of the sequence, evaluates the likelihood of a pixel color value to belong to the background. With every pixel that matches the background the pixel model is updated, adaptively learning smooth illumination changes:

$$\mu_t = (1-\rho)\mu_{t-1} + \rho\mathbf{I_t} \quad (11)$$

$$\mathbf{\Sigma}_t = (1-\rho)\mathbf{\Sigma}_{t-1} + \rho(\mathbf{I_t} - \mu_{t-1})^T(\mathbf{I_t} - \mu_{t-1}) \quad (12)$$

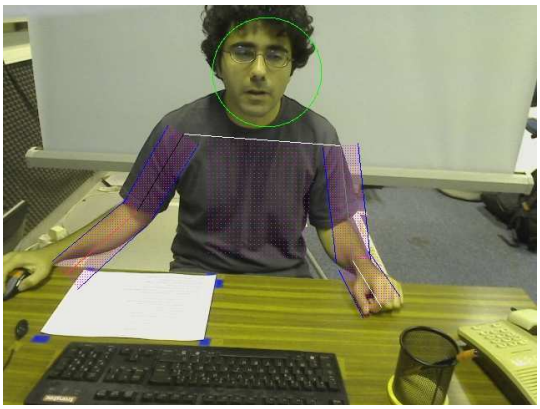
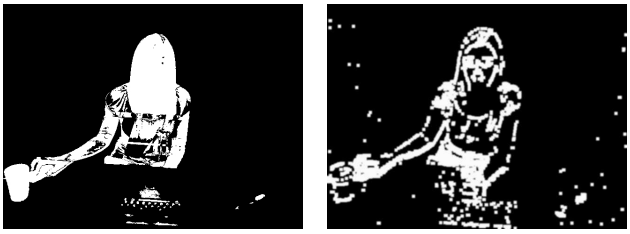


Fig. 3. Projection of the flesh model associated to a given particle

A shadow removal algorithm [65], based on the color and brightness distortion, is used to enhance the segmentation.



(a) Foreground Mask (b) Contours Mask

Fig. 4. Extracted Image features

Contour detection is performed by means of the Canny edge detector [10]. The result is dilated with a 8-connectivity and 5x5 structuring element, and smoothed with a Gaussian mask. In order to avoid spurious contours, we subtract the background contours. This implies also deleting some pixels in the edges of interest but the body structure it's in general preserved. Finally, a simple skin detection method based on evaluating the likelihood ratio between skin and non-skin hypothesis is performed. The likelihood functions are estimated by 8-bins color histograms of several skin and non-skin samples.

The likelihood evaluation procedure involves the projection of the flesh model of every particle onto the image coordinate system. The resulting shape is scanned and matched with the foreground segmentation. The weight is computed as follows:

$$\omega^{fl} = \frac{1}{N}\sum_{n=1}^{N}(1 - I_n^f) \tag{13}$$

Since pixel intensities in the foreground masks (I_i^f) have 0 or 1 as possible values, the weighting function is obtained by a normalized sum of the background pixels falling inside the projected flesh model. In the head model case, we add skin detection information:

$$\omega^{fh} = \frac{1}{N}\sum_{n=1}^{N}(1 - I_n^f I_n^s) \tag{14}$$

Therefore, the final foreground weight ω^f is the averaged sum of all the limbs and head weights. Foreground segmentation provides data that are generally invariant to clothing and most of the background conditions.

Since many configurations can be explained via this feature, foreground information is used to penalize false poses rather than to single out the correct one. Moreover, the proposed measure shows how well the model fits the observation, but doesn't evaluate how well the observations are being explained by the model. Suppose the likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$ is available and that a given pose generates a pdf. A measure that can be used to assess the similarity of the likelihood and the generated pdf is the Kullback-Leibler divergence. At this point is important to remark that the KL divergence will provide different results depending on the factor order (except if both pdfs are identical). We can establish an analogy with our likelihood approximation. We are trying to determine the mutual information of the model and the observations. Therefore, we propose to include an additional divergence measure between the projection of the flesh model and the foreground masks to see how well a particle explains the observations.

$$\omega^d = \frac{1}{N_f}\sum_{n=1}^{N_f}(I_n^f(1 - B_n)) \tag{15}$$

This divergence basically aims for projecting a given particle and computing the overlap between the pixels B_n of this projection and the N_f foreground pixels of the observation.

Contours found in the body usually provide good information on the location of the arms and the legs. However, in some cases, clothing and background can introduce spurious contours that reduce the reliability of this feature. As mentioned above, we try to minimize the background impact by subtracting the background contours. The proposed weighting function for this feature is a sum of squared differences between the contour pixels and the edges of the flesh model aligned with the axis of the limb:

$$\omega^e = \frac{1}{N}\sum_{n=1}^{N}(1 - I_n^e)^2 \tag{16}$$

Finally, all these weights are combined for every camera:

$$\omega = exp\left(\sum_{c=1}^{C}(\lambda_c^f\omega^f + \lambda_c^e\omega^e + \lambda_c^d\omega^d)\right) \tag{17}$$

We use a set of weights for every camera and measure to adjust the importance of every feature according to its importance and visibility. Since in our scenario the subject stays in his seat, we assume that the visibility component can be determined beforehand.

D. Annealing Particle Filter

It has been shown in several works that SIR Particle Filters are a good approach for tracking in low dimensional spaces, but they become inefficient in high-dimensional problems. Deutscher et. al [15] proposed a variation of the SIR framework by introducing the concept of Annealing PF. In body pose tracking problems, the likelihood approximation often is a function which has several peaked local maxima. Annealing PF deals with this problem by evaluating the particles in several smoothed versions of the likelihood approximation. After the weights are computed via the modified likelihood, particles are resampled and propagated with Gaussian noise with zero mean and a covariance that decreases at every step. Each one of this steps (weighting with a smoothed function, resampling and propagation) is called an annealing run. In the last annealing run the estimation is given by means of the Monte-Carlo approximation of the posterior mean:

$$\hat{\mathbf{x}}_\mathbf{t} = \sum_{i=1}^{N_s} w_t^i \mathbf{x}_t^i \tag{18}$$

The most usual way to smooth the weighting function is by means of an annealing rate, an exponent $\beta < 1$. In the first layer β is minimum but it progressively increases with each layer, sharpening the likelihood approximation. In [15] a method for tuning β with the survival rate after each annealing run is proposed.

The sharpness of the likelihood function is due to the high dimensional space in which is defined, the use of a hierarchical model [11] is another possible strategy in order to have annealing layers. Since our model is quite simple, a hierarchical approach is not justified. We have implemented an annealing particle filter in which the smoothing is done by means of an exponent β. In our case, the annealing rate is updated according to the survival rate of the preceding layer $\alpha(\beta_{t-1})$. Given a desired survival rate α_T:

$$\beta_t = \beta_{t-1} - \lambda(\alpha_T - \alpha(\beta_{t-1})) \tag{19}$$

Due to the image feature characteristics, we also introduce β in (17), giving higher importance to the foreground-based measures in the first layers and to the contour-based measures in the last layers.

$$\omega = exp\left(\sum_{c=1}^{C}(\lambda_c^f(\frac{1}{\beta})\omega^f + \lambda_c^e(\beta)\omega^e + \lambda_c^d(\frac{1}{\beta})\omega^d)\right) \tag{20}$$

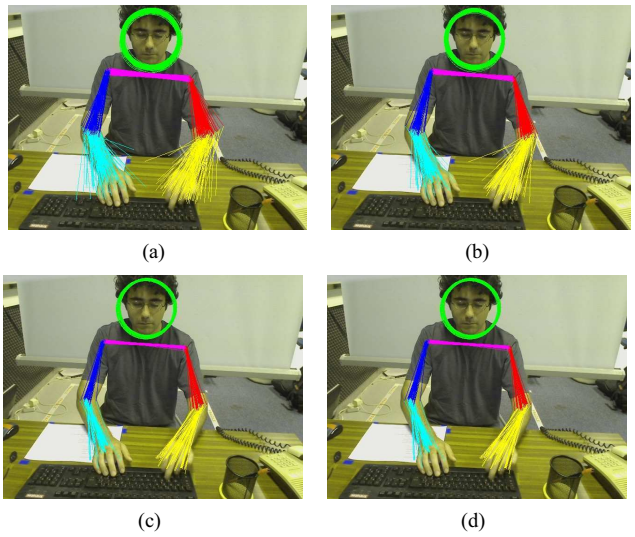


Fig. 5. Annealing Layers. The covariance used in the propagation step is progressively reduced through four annealing layers while the estimator gets closer to the true pose

Therefore, we propose to work with overall smoothing and feature-based smoothing. However, more work needs to be done in this area in order to show that this approach can help to efficiently reach the true pose.

IV. Sound-Based Event Detection

This section deals with detection of sound activity and classification of sounds into the typical events that would be encountered. In the first step, any sort of sound activity is detected and in the second step it is classified. The details of each step are explained below.

A. Sound Activity Detection

The field of Sound Activity Detection has been researched for several years. Most of the research has been in the field of Voice activity detection in noisy conditions. This is essentially different from the current experiment in which all sound activity needs to be detected. This makes it a slightly difficult problem in a way, because a threshold on the length of the activity cannot be provided. The detection has to be made on short bursts of sounds like clicks of mouse as well as continuous speech. So a dynamic threshold needs to be provided, based on the current noise level.

Previous work done in voice activity detection was mainly by Mak *et al.* [38] and Nemer *et al.* [45]. Nemer *et al.* proposed a method based on the residual of the signal, and used higher order statistics of the noise in order to set the threshold to detect sound activities. Renevey and Drygajlo [52], proposed an Entropy based threshold for activity detection. The method used in this experiment uses the entropy found on the residual as a measure to detect activity.

The following steps are taken to detect sound activity

- **The signal is windowed with a window size of 40 ms and a shift of 20 ms.**
- **The signal within one window is approximated by 2 Linear Prediction Coefficients (LPC). This is done to grossly approximate the frequency spectrum and calculate the bias.**
- **The residual of the signal, which is the error between the estimated LPC and the true signal is calculated. Fig. 6 shows the spectrum of the signal and Fig. 7 shows the corresponding residual. One can observe that the bias has been canceled and the spectrum has been whitened.**
- **The Entropy is calculated for the residual, assuming a gaussian distribution, since whitening has been performed. The evidence of activity is given by the entropy. A higher entropy indicates a higher level of activity.**
- **A dynamic threshold is calculated, which decides whether the entropy is high enough to be classified as noise.**

The biggest problem with sound activity detection is the hysteresis associated with detection. After detecting a certain sound, we cannot hear other less louder sounds occurring after it. Hence a dynamic threshold has to be calculated based on the statistics of the past. Since the distribution of the sound activity entropy is unknown, a histogram is calculated, for the entropy over a history of around 10 seconds. If the entropy level is in the highest L% range of the histogram, then it considered as activity. However, the entropy level has to go below the 50% range of the past activity to be classified as background noise. Fig. 8 shows the Entropy variation of a short segment of the signal. The two dynamic thresholds are also indicated along with the decision.

The value of L decides the region in the Detection Error Trade-off (DET) curve as shown in Fig. 9. Most of the errors that occur are due to the fact that length of the detected activity is either shorter or longer than the annotated activity. Often what is annotated as a contiguous activity is split into several activities or what is annotated as different activities, is detected as a single activity. The DET curve for length independent detection is shown in Fig. 10.

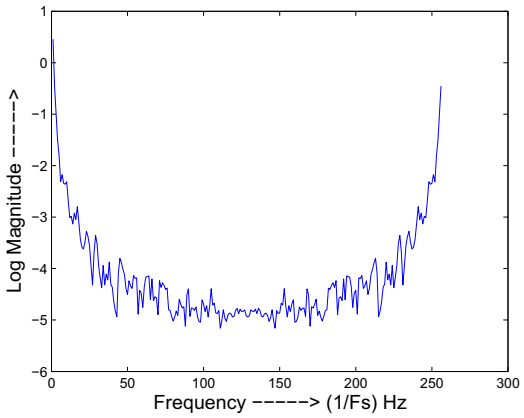


Fig. 6. The log-frequency spectrum of a typical signal

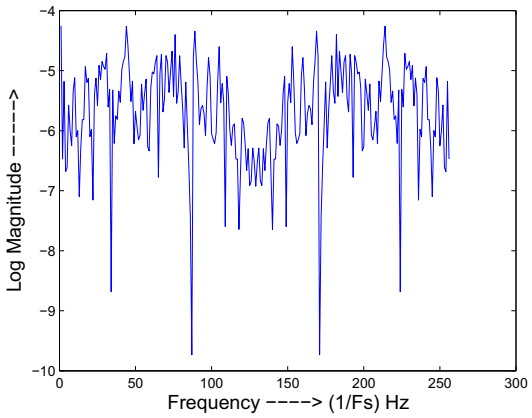


Fig. 7. The log-frequency spectrum of the residual

B. Sound Event Classification

Sound event classification has been commonly called auditory scene analysis in literature. The most seminal work on auditory scene analysis is discussed by Bregman [8]. Several methods and several features have been tried for this purpose. Among the most common features used are Bark-filter coefficients, wavelet coefficients, Linear Prediction Coefficients etc. Similarly, Support Vector Machines (SVM), Self Organizing Maps (SOM), Artificial Neural Networks (ANN) and Gaussian Mixture Models (GMM) and their combinations have been used for this purpose.

In our experiments Bark filter coefficients are used as features for classification, because the Bark filters mimic the subjective measurements

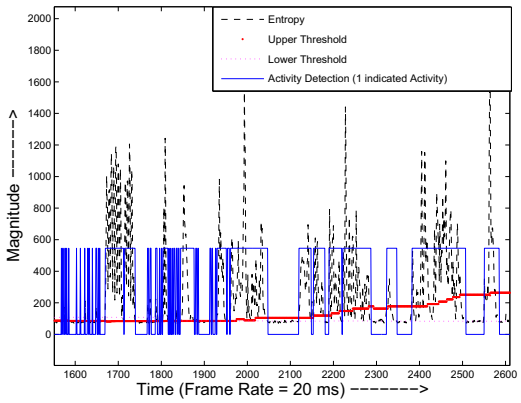


Fig. 8. The Entropy variation for a short segment of the signal. The dynamic thresholds are also shown

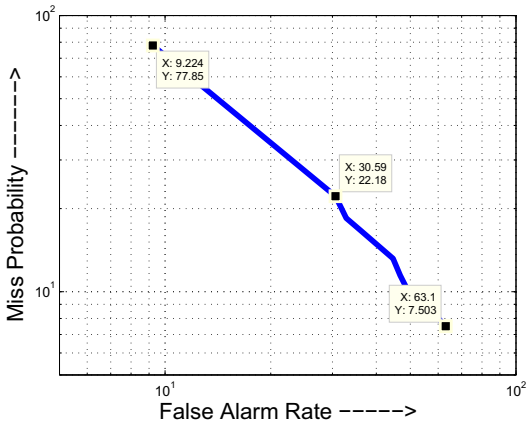


Fig. 9. The log scale plot of the Detection Error Trade-off Curve

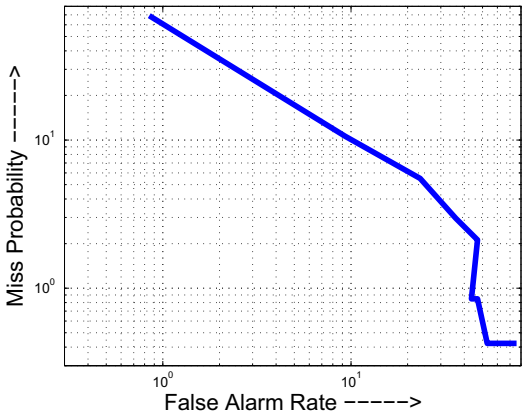


Fig. 10. The log scale plot of the Detection Error Trade-off Curve, independent of length

TABLE II
SOUND EVENT RECOGNITION RESULTS

Sound	Accuracy	False Alarm	Most confused
Voice	67.3%	12.3%	Pencil keep
Telephone ringing	54.3%	0%	Voice
Writing sound	5.2%	0%	Silence
Keyboard typing	45.0%	15.3%	Mouse click
Glass use	89.3%	56.3%	-
Mouse click	43.3%	19.5%	Typing
Phone receiver	63.3%	32.4%	Keyboard Typing
Pencil use	54.3%	12.8%	Voice
Overall	53.8%	23.7%	-

of loudness of the human ear. Since we have sound events with different durations, and since we are classifying contiguous blocks of signals, one-state HMMs are used for event classification where the observation probability distribution is expressed with a GMM. This helps in coupling the likelihoods of each of the frames of the signal to give a single likelihood value.

The most important question in these models is to decide how many mixture components will be employed. This is a difficult problem, especially because there are only a few available sounds, with varying length and duration. The number of Gaussians for each sound class is decided by maximizing the Bayesian Information Criterion (BIC). The sound classes that we used are as follows:

1) **Voice**
2) **Telephone Ringing**
3) **Typing Sound**
4) **Writing sound (with a pencil)**
5) **Placing the glass on the table**
6) **Clicking of the Mouse**
7) **Picking up the phone receiver or putting it back**
8) **Picking up or placing the pencil**

One can see that, a few of these sounds are quite similar and difficult to distinguish even for human beings. However, since the experiment is set-up in a controlled environment, one can expect a decent performance. Table II denotes the results of the recognition of each of the sounds in the list.

As we can see, the accuracy is higher for detection of voice and glass use, but the false alarm is also high for the same two sounds. There is a very high confusion rate between mouse click and typing for example. It is expected that we assume higher priors for more probable events and lower priors for less probable events. However in that case most of the sounds would be classified as voice, because the voice includes sounds similar to each of these mentioned sounds. So the classification is done assuming an equal prior. The overall accuracy may be boosted if the priors were selected according to their probability of occurrence, but then the overall accuracy evaluation would be biased. It does not make sense to use a weighted average for calculation of accuracy, because one wrongly classified event with low probability would affect the overall accuracy greatly.

More work can be done in the direction of a better classifier, using combination of GMM with classifiers like ANNs or SVM. More evaluation is necessary to deal with different time lengths of each of these sound events. Different number of models and modeling the dynamics of the sounds could be other options. A varying length window in order to calculate the Bark coefficients maybe another direction of research.

V. SPEAKER VERIFICATION

The speaker verification system provides a Boolean authentication decision based on the analysis of a speech fragment. Speech-based verification systems can be classified into two main types. In the first approach, the speaker utters a word or a sentence, which is fixed for all authentication attempts. This is called the text-dependent approach. In the more difficult text-independent approach, which is more appropriate for this scenario, the speaker can utter any sentence, and the textual content is not known a priori. For a good survey of speaker verification systems, the reader is referred to [47]. Suffice it to say that all such systems need a speaker model, and an impostor model to determine the decision for authentication. Frequently employed methods for modeling the speaker as well as the impostor include dynamic time warping (DTW),

vector quantization (VQ), Gaussian mixture models (GMM), and hidden Markov models (HMM).

The DTW is used for non-linear aligning of two time sequences and computing the minimum distance between them. Use of DTW in speaker verification system is based on assumption that every speaker is uttering the same word or sentence approximately in the same manner but differently from other speakers. Here the speaker is represented by a template of one or limited set of words or sentences. As such, this method is not adequate for text-independent verification. Vector quantization methods are based on the assumption that the acoustic space of a speaker's speech output can be divided into non-overlapping classes, representing different kinds of sounds, for example phonemes. Each class is defined by one vector, centroid and so each speaker is represented by a set of these classes, thus by his own codebook of vectors. In the GMM approach, the codebook vectors are the means of the Gaussian distributions. Here, the noise around each mean is assumed to be normally distributed. Each speaker is represented by a Gaussian mixture density, which is weighted linear combination of Gaussian distributions of each speaker's acoustic class. Thus speaker is represented by a set of weigths, means and variances. In the HMM approach, the speech dynamic is modeled by a Markov model, where the states are modeled by codebooks of the VQ (discrete HMM) or by Gaussian mixture densities (continuous and semi-continuous HMM). In the particular case of text-independent verification systems, ergodic models are preferred where all interstate-transitions have non-zero probabilities.

In this work, we follow the GMM approach based on the results reported in [4], [51]. First, a number of features are extracted from the input signal. Following [20], we use Mel-filter cepstral coefficients (MFCC) by applying the following transformations:

- Preemphasis filter
- Division of signal into frames
- Fast Fourier transformation for obtaining frequency spectrum
- Logarithmic transform
- Application of Mel-filter banks to the spectrum
- Discrete cosine transform

In speech recognition, usually 13 coefficients are selected from the MFCC. The first and second derivatives (i.e. velocity and acceleration) are added to these coefficients to indicate the history and evolution of the signal, resulting in 39-dimensional feature vectors. N-dimensional feature vector implies using N-dimensional Gaussian distributions, thus the N-dimensional mean and NxN covariance matrix. Because of sufficient effectiveness in modeling the components are restricted to have diagonal covariances.

Once a speaker model is learned, there are two ways of authenticating a particular speaker [47]. In the first approach, a threshold is selected for the probability $P(\lambda^t|O)$, where λ denotes the model parameters of the target speaker, and O is the observed signal. In the second approach is a treshold selected to the ratio of proability of the genuine speaker and probability of the impostor model, which is trained on all speakers in the system except the genuine speaker class. This implies that for every person in the system, two models will be trained. In the case of sufficiently many subjects, a single and generic impostor model can be employed. The implementation of the GMM approach is done by using the Hidden Markov Models Toolkit (HTK) [66]. The GMM was built as an HMM with just one state, as shown in Fig. 11.

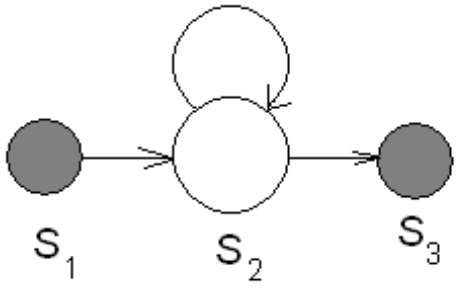


Fig. 11. One-state HMM

VI. Contact-Based Biometrics

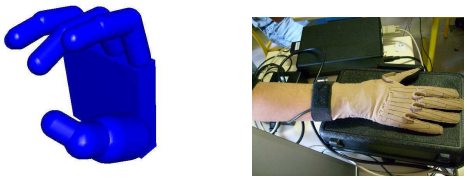
The concept of Contact-Based Biometrics derives from the simple observation that every person handles the objects of the surrounding environment quite differently. For example, the action of picking up a glass or holding a knife depends on the physiological characteristics of each person and the way that this person is used to manipulate objects. Contact-Based Biometrics can also been thought as a specialized part of Activity-Related Biometrics for every activity which involves an object.

In the context of this project we intend to investigate the feasibility of such biometric features in user authentication applications. The proposed approach exploits methods from different scientific fields, such as collision detection and pattern classification, to solve the problem of authentication. The major parts of the final implementation scheme are the setup of a 3D virtual environment, the registration of the user and the objects in this environment, the extraction of collision features during an action between the user and an object and the classification procedure.

A. 3D Environment Setup and Model Registration

Collision detection algorithms can only be used in a 3D environment with full knowledge of the geometry of each object. The virtual environment of the presented pilot requires only the 3D representation of the user's hand and each object that is of interest. The user's hand is modeled as a set of five fingers connected to the palm, which is modeled as a simple rectangle (Figure 12(a)). Each finger has four degrees of freedom (DOF) and consists of three phalanxes which are modeled as simple capsules.

For the registration of the hand we used the CyberGlove® (http://www.immersion.com/3d/). The CyberGlove® (Figure 12(b)) provides the angles between the phalanxes of the hand, so it is possible to reconstruct the 3D representation of the hand. Note, that the virtual representation of the hand is not perfectly accurate because the size of the fingers and the phalanxes are not known. In order to satisfy the requirements of a realistic pilot we cannot make any assumptions or measures on the user so this inaccuracy is considered as noise.



(a) (b)

Fig. 12. (a) The 3D representation of the hand. (b) The CyberGlove®.

The objects of the environment can be registered using computer vision techniques for object tracking. However, it is not absolutely necessary to have an accurate representation of the object in the virtual environment. In particular for rigid objects, which are typically encountered in an office environment, we can simplify the geometry of the object using a priori information. This simplification is possible as the real shape of each object is mostly related to the specific action that is used and not to the way it is handled. For example, a glass can be represented by a cylinder since the user grabs only the outer surface of the glass.

B. Collision Feature Extraction

The classification features consist of any information that can be acquired by employing state-of-the-art algorithms for proximity queries. These include penetration depth [24], closest distance [26], [32], contact points etc. The literature in the field is vast and there are numerous algorithms to accurately perform queries in real-time. The interested reader is directed to [17], [33], [57], [58], [60] for further details. For our purposes we used the algorithms for rigid convex objects [59], [60] of the software package SOLID (http://www.dtecta.com/).

Proximity queries are performed between the object and every finger of the user's hand. Each query refers to either of two states, collision or no collision between the two virtual shapes. For example penetration depth can only be calculated when two objects intersect since it is always zero otherwise. However, in a user-object interaction scheme it is necessary to continuously produce discriminant feature samples. Thus, any proximity query as a single feature would not provide adequate information to a classifier.

In the proposed method we employ the combination of the penetration depth and the closest distance, depending on the collision state, to define the feature space. The penetration depth and the closest distance are usually described as 3D vectors in virtual simulations. However, in our case we prefer to describe them as the pair of points (p_{finger}, p_{object}), one on the finger and the other one on the object, that define the respective vector $\mathbf{v} = p_{finger} - p_{object}$. This way the 3D position of each finger affects the values of the feature vector, while $\mathbf{v}$ would only describe the relative direction which is most probable to be similar even for different fingers. Let pd_k and cd_k denote the points of the penetration depth and the closest distance respectively for either the finger or the object k. The feature sample $f_e(i, O)$ for the finger e and the object O on the i-th frame is

$$f_e(i,O) = \begin{cases} (pd_e, pd_O), & \text{e and O collide} \\ (cd_e, cd_O), & \text{e and O do not collide} \end{cases}$$

The final feature vector $F = \bigcup_e \{f_e\}$ is formed using the collision information from all the five fingers and is a 30-dimensional vector.

VII. Continuous Face Authentication

With the rapid increase of video surveillance equipment and webcam usage, it became necessary to develop robust recognition algorithms that are able to recognize people using video sequences, which not only provide abundant data for pixel-based techniques, but also record the temporal information. This project inspects two complementary approaches to face biometrics from continuous video, detailed in this section and the next.

The processing for the face and facial motion analysis modules starts with detecting the face. We use the OpenCV face detection module that relies on the adaboosted cascade of Haar features, i.e. the Viola-Jones algorithm for this purpose [61]. The face camera is positioned so that the face image roughly covers a 150×150 pixel area, which changes greatly as the subject moves around.

One of the assumptions we have in the face authentication module is that the statistical models that incorporate general face information are trained offline, prior to the actual experimental setup. This means that the bulk of the training database should consist of external data. For this purpose, we have used the world model of 300 face images that accompany the BANCA database [3], enriched with one gallery image per enrolled person. This is a realistic assumption, and since the gallery is acquired with different illumination conditions as well, the actual experimental environment presents a formidable challenge, with completely uncontrolled illumination under ordinary (and poor) office lighting.

For continuous face authentication, we take a straightforward approach. The detected faces are cropped, rescaled to a fixed size, projected to a previously computed subspace, and compared to the templates residing in the gallery. For controlling the illumination, we apply a image enhancement procedure proposed by Savvides and Kumar [54]. In this procedure, the pixel intensities are mapped to a logarithmic range, which nonlinearly allocates a broader range to dark intensity levels, increasing the visibility.

The subspace is found by applying the Karhunen-Loeve transform to the enhanced training set. The matching of a claim with a gallery image can be achieved by thresholding a Mahalanobis-cosine distance between projected vectors. If the subspace-projected query is denoted by $\boldsymbol{u} = [u_1 u_2 \dots u_p]'$ and the subspace-projected gallery template is denoted by $\boldsymbol{v} = [v_1 v_2 \dots v_p]'$, denote their corresponding vectors in the Mahalanobis space with unit variance along each dimension as:

$$m_i = \frac{u_i}{\sigma_i} \tag{21}$$

and

$$n_i = \frac{v_i}{\sigma_i} \tag{22}$$

where σ_i is the standard deviation for the i^{th} dimension of the p-dimensional eigenspace. Then the Mahalanobis cosine distance is given by [49]:

$$d_{MC}(\boldsymbol{u}, \boldsymbol{v}) = cos(\theta_{mn}) = \frac{mn}{|m||n|} \tag{23}$$

A. Adaptive Cropping

The preprocessing of the external database is not replicated in our acquisition conditions. This means that the eigenspace projection that models the variation in aligned face images is not necessarily the ideal projection for a given query image. To remedy this situation, we apply an adaptive cropping algorithm that fine-tunes the face detection result so as to minimize the reprojection error e. Assume the eigenspace is denoted with $[\lambda, \boldsymbol{e}]$, where λ stands for the sorted eigenvalues and $\boldsymbol{e}$ are the corresponding eigenvectors. The projection of query $\boldsymbol{x}$ to the eigenspace is:

$$\boldsymbol{u}_{(p\times 1)} = \boldsymbol{e}'_{(p\times d)}(\boldsymbol{x}_{(d\times 1)} - \mu_{(d\times 1)}), \tag{24}$$

where μ denotes the data mean, and the subscripts indicate dimensionality. The reprojection error is given by:

$$e = ||\boldsymbol{x}_{(d\times 1)} - \boldsymbol{e}_{(d\times p)}\boldsymbol{u}_{(p\times 1)} + \mu_{(d\times 1)}||. \tag{25}$$

The pseudocode of the algorithm is given in Fig. 13. Fig. 14 shows the cumulative effect of illumination correction and adaptive cropping on a sample frame.

```
algorithm Adaptive_Cropping(faceImg)
    cropping ← [0,0,0,0]
    oldError ← Infinity
    found = False
    cropDir = 1
    while NOT found
        oldError ← newError
    /*Crop the image in one of four directions*/
        cropping(cropDir) ← cropping(cropDir) + 1
        croppedImg ← crop(faceImg,cropping)
    /*Scale to fixed size*/
        scaledImg ← scale(croppedImg)
    /*Illumination normalization*/
        normalizedImg ← logTransform(scaledImg)
    /*Projection*/
        projImg ← eigenVectors'*(normalizedImg-meanImg)
    /*Re-projection into the original space*/
        reprojImg ← (eigenVectors*projImg)+meanImg
    /*Update the error*/
        reprojError = norm(reprojImg - normalizedImg)
        if reprojError < oldError
            newError ← reprojError
        else
    /*Reverse the cropping*/
            cropping(cropDir) ← cropping(cropDir) - 1
        end
    /*Update the next cropping direction*/
        cropDirection ← mod(cropDirection+1,4)
        found ← (updated in the last cycle of four directions)
    end
    return cropping
end
```

Fig. 13. Adaptive Cropping Algorithm

Fig. 14. a) The original captured frame. b) The illumination compensated image. c) The result of the adaptive cropping

B. Probabilistic Matching

The activity model necessitates a short video sequence to be recorded for training purposes. This allows us to use a larger training set for the face authentication module as well. For each subject in the gallery, one sequence of recordings is processed with the face detection and adaptive

cropping modules. The ensuing cropped images are projected to the Mahalanobis space, and modeled with a mixture distribution.

The general expression for a *mixture model* is written as

$$p(\boldsymbol{x}) = \sum_{j=1}^{J} p(\boldsymbol{x}|\mathcal{G}_j)P(\mathcal{G}_j) \tag{26}$$

where $\mathcal{G}_j$ stand for the components, $P(\mathcal{G}_j)$ is the prior probability, and $p(\boldsymbol{x}|\mathcal{G}_j)$ is the probability that the data point is generated by component j. In a *mixture of Gaussians* (MoG), the components in Eq. 26 are Gaussian distributions:

$$p(x|\mathcal{G}_j) \sim \mathcal{N}(\mu_j, \boldsymbol{\Sigma}_j) \tag{27}$$

Typically, the covariance expression is restricted in MoG models to control the complexity of the model, as a diagonal covariance scales linearly with dimensionality, whereas a full covariance scales quadratically. In this work we use the factor analysis approach to model the covariance, where the high dimensional data $\boldsymbol{x}$ are assumed to be generated in a low-dimensional manifold, represented by latent variables $\boldsymbol{z}$. The *factor space* spanned by the latent variables is similar to the principal space in the PCA method, and the relationship is characterized by a *factor loading matrix* Λ, and independent Gaussian noise $\boldsymbol{\epsilon}$:

$$\boldsymbol{x} - \boldsymbol{\mu}_j = \boldsymbol{\Lambda}_j \boldsymbol{z} + \boldsymbol{\epsilon}_j \tag{28}$$

The covariance matrix in the d-dimensional space is then represented by $\boldsymbol{\Sigma}_j = \boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j^T + \Psi$, where Ψ is a diagonal matrix and $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, \Psi)$ is the Gaussian noise. We obtain a *mixture of factor analysers* (MoFA) by replacing the Gaussian distribution in Eq. 26 with its FA formulation.

To learn the distribution of training faces of a single class, we use the incremental mixtures of factor analysers (IMoFA) algorithm, which automatically determines the number of components in the mixture, and tunes the latent variable dimensionality for each mixture component separately. For more details, the reader is referred to [53]. The ensuing model for the subject is $(\boldsymbol{\Lambda}_j, \boldsymbol{\mu}_j, \boldsymbol{\epsilon}_j, \pi_j)$, with π_j being the component prior, and j is the index for mixture components. The authentication of a normalized and projected image $\boldsymbol{x}_t$ is effected by checking a pre-fixed threshold:

$$p(x_t|\mathcal{G}) \geq \tau \tag{29}$$

At any point in time, the continuous face authentication module evaluates the most recent frame, and returns a Boolean decision. The threshold τ depends on the Mahalanobis space dimensionality, and scales approximately linearly with it. For a 300-dimensional Mahalanobis space, we have used a threshold of -400 for the log-likelihood, a higher value will reject more frames and ensure a more secure system, whereas a lower value will favour user convenience over security. It is also possible to base the decision on all the frames up to time t, by using any classifier combination method.

VIII. Behavioral Face Biometrics

The previous section dealt with the static facial appearance, ignoring the behavioral cues that can be potentially useful for discriminating identities. Recently there is much attention to biometric systems that exploit temporal information in videos, and most of the proposed approaches involve a heterogeneous mixture of techniques. These approaches can roughly be classified into the following categories:

- Holistic approach: This family of techniques analyze the head as a whole, by extracting the head displacements or the pose evolution. In [30] Li et al. propose a model-based approach for dynamic object verification and identification using videos. In 2002, Li and Chellappa were the first to develop a generic approach for simultaneous object tracking and verification in video data, using posterior probability density estimation through sequential Monte Carlo methods [29]. Huang and Trivedi in [19] describe a multi-camera system for intelligent rooms, combining PCA based subspace feature analysis with Hidden Markov Models (HMM). Liu and Cheng proposed a recognition system based on adaptive HMMs [35]. They first compute low-dimensional feature vectors from the individual video frames by applying a Principal Component Analysis (PCA); next they model the statistics of the sequences and the temporal dynamics using a HMM for each subject. In [1] Aggarwal et al. have modeled the moving face as a linear dynamical system using an autoregressive and moving average (ARMA) model. The parameters of the ARMA model are estimated for the entire database using the closed form solution. Recently, Lee et al. developed a unified framework for tracking and recognition, based on the concept of appearance manifold [28]. In this approach, the tracking and recognition components are tightly coupled: they share the same appearance model.
- Feature based approach: The second group of methods exploits the individual facial features, like the eyes, nose, mouth and eyebrows. One of the first attempts to exploit facial motion for identifying people is presented by Chen et al. in [12]. In their work, they propose to use the optical flow extracted from the motion of the face for creating a feature vector used for identification.
- Hybrid approach: These techniques use both holistic and local features. Colmenarez et al. in [14] have proposed a Bayesian framework which combines face recognition and facial expression recognition to improve results; it finds the face model and expression that maximizes the likelihood of the test image.

This section proposes a new person recognition system based on temporal features from facial video. As in the previous section the face area is first detected in each frame of the video. The registration, or the alignment problem, however, has different criteria to satisfy. Since we will track the features, the alignment is not absolute, but relative to the previous frame, minimizing a mean square error measure. For aligned faces, the optical flow is calculated from consecutive frames, and used as feature vectors for person recognition.

Once the faces are detected with the Viola-Jones method, a representation called the "integral image" is created using Haar-like features.

The learning algorithm is based on AdaBoost, which can efficiently select a small number of critical visual features from a larger set, thus increasing performance considerably.

Next the resulting image is cropped as shown in Fig. 15 based on anthropological measures to limit the image to facial features that exhibit more motion.

Fig. 15. Detected and cropped face images in two frames.

Face alignment was required due to the simple fact that we wanted to focus our attention on motion of local features from the face such as the lips and the eyes. If this step is not performed before feature extraction, global motion of the head significantly affects the results. Alignment of the faces detected in two different frames was carried out by minimizing the mean square error of the integral image difference:

$$\arg\min \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (I_1(i,j) - I_2(i,j))^2 \tag{30}$$

where,... Fig. 16 shows two facial images found in consecutive frames aligned with this method.

Fig. 16. Two facial images aligned and superimposed.

We have decided to use optical flow vectors for person recognition, calculated by the Lucas-Kanade technique [36], which uses the spatial intensity gradient of the images to guide the search for matching locations, thus requiring much less comparisons with respect to algorithms that use a predefined search pattern or search exhaustively. Then block means are

taken to reduce the size of the feature vector to standard dimensionality of 200. Fig. 17 shows the optical flow computed from the images aligned in Fig. 16.

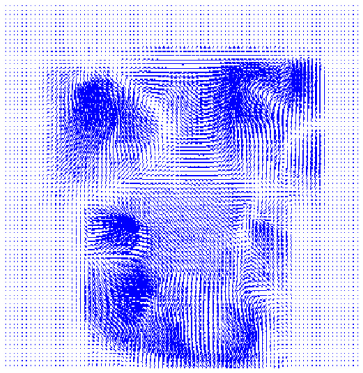


Fig. 17. Optical flow from consecutive frames.

IX. Continuous Activity - Related Biometric Authentication

Among the project's prominent objectives is to investigate the effectiveness and applicability of activity - related biometric technologies. Activity - related biometrics is a novel and innovative concept in biometric user authentication and refers to biometric signatures extracted by analyzing the response of the user to specific stimuli, while performing predefined but natural work - related activities. The novelty of the approach lies in the employment of dynamic features extracted by the moving human model as biometric signal, as well as in the fact that the biometric measurements will correspond to the user's response to specific events, being, however, fully unobtrusive and fully integrated in the user's workspace. The activity - related biometric authentication module evaluates the fundamental assumption that each user's dynamic behavioral profile contains unique intrinsic characteristics that can be used for authentication. Furthermore, a reliable implementation of an activity - related biometric authentication system is ideal for continuous user authentication, thus alleviating the main limitation of some successful state-of-the-art approaches (fingerprint, iris etc.) which cannot be recovered once forged.

In the following, the modules and methods that were implemented to perform activity - related authentication will be described. In addition to that, the pilot setup and the experimental procedures followed in order to evaluate activity - related biometrics will be presented.

A. Activity Detection and Recognition Module

As stated above, the user's dynamic profile extraction is based on the response to specific environment - generated stimuli. Any human behavior is associated to some action or activity. The aim of stimuli generation is to trigger the execution of specific actions by the user, upon which his behavioral profile can be then calculated. It is therefore clear that the extraction of the activity - related features must be preceded by an action detection, segmentation and recognition procedure. This goal is achieved by means of a multimodal approach that uses the output of the sound event recognition Module, the Object Occlusion Tracking Module and the body motion tracking module along with a Coupled Hidden Markov Model formulation in order to detect the generation of the stimuli and segment the user's response (action). The segmentation output of the Activity Recognition Module can be then fed to the Activity - Related Biometric Authentication Module. Fig. 18 illustrates the above inter - module relationships.

Numerous relevant approaches for activity recognition have been reported in the literature using object manipulation context information [43], [46], [64] and/or object trajectory information in the given scene [5], [31]. Sound event detection has also been previously employed to assist inference of ongoing activities [55], [63].

The proposed method for Activity Recognition is based on the detection of three different kinds of Scene Events occurring in the scene: Sound Events (e.g. Phone Ringing), detected by the sound event recognition Module, Proximity Events (e.g. "Hand close to Glass"), detected by the

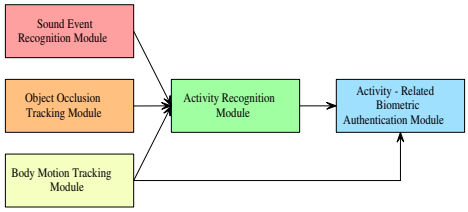


Fig. 18. Module cooperation for Activity Recognition

Human Body Tracking Module along with predefined knowledge of the object positions on the controlled workspace and Object Occlusion Events detected by the respective tracker. An Object Occlusion Event is emitted when some object in the scene is missing from its "normal" position.

In order to achieve action recognition, a two - stream Coupled HMM is associated to every action class and trained on two sets of discrete observation symbols (one for each stream) extracted by the primitive events described above (i.e. second layer events). The first set of second layer symbols is a subset of the Sound Event set that can be associated to a particular action. For example, the Phone Conversation Coupled HMM only handles relative sound events (Ringing, Speech, Silence etc.) and disregards the rest (e.g. Writing sound). The observation symbols of the second stream are formed as meaningful (for the particular action class) combinations of the Object Occlusion and Proximity Events of the first layer. For instance, the state "Phone receiver missing" AND "Left Hand close to Head" forms a single second layer event that is used as observation symbol of the second stream of the Phone Conversation CHMM to represent the state of "talking on the phone".

At every timestamp of some activity sequence, first and second layer events are detected and form N double - stream discrete observation sequences, where N the number of actions to be recognized and segmented. Each CHMM uses an overlapping sliding window that goes through its own observation sequence. The size of the sliding windows and the size of overlapping are experimentally defined. The CHMM of each action is trained on manually annotated sequences and a probability threshold is defined, above which the respective action is recognized and a portion in the size of the sliding window is segmented and fed to the Activity Biometrics Module. Fig. 20 graphically depicts the Activity Recognition Module. The reason for performing the mapping from first layer events to second layer events is to impose a smaller size on the final observation sets and process the three initial streams of events into only two - stream Coupled HMMs, which results in making training more efficient.

B. Coupled Hidden Markov Models

The need of a Coupled Hidden Markov Model formulation is justified by the fact that Scene Event detection is often erroneous, producing many false alarms, wrong inferences and multiple occlusions over time. Consequently, detected event symbols would better be thought of as the probabilistic output of some underlying process, rather than as deterministic events. Furthermore, Coupled HMMs offer a robust mathematical background for integrating multimodal observations and fusing different but correlated processes (sound events + human activity based events).

Our Coupled HMM implementation is based on the formulation presented by Ara V. Nefian et al. [44], where the hidden nodes of each stream interact and at the same time have their own observations (Fig. 21). The elements of the CHMM (Initial, Transition and Observation probabilities) are described as:

$$\pi(\mathbf{i}) = \prod_s \pi^s(i_s) = \prod_s P(q_1^s = i_s) \tag{31}$$

$$b_t(\mathbf{i}) = \prod_s b_t^s(i_s) = \prod_s P(\mathbf{O}_t^s | q_t^s = i_s) \tag{32}$$

$$\alpha(\mathbf{i}|\mathbf{j}) = \prod_s \alpha^s(i_s|\mathbf{j}) = \prod_s P(q_t^s = i_s | \mathbf{q_{t-1}} = \mathbf{j}) \tag{33}$$

The CHMMs are trained using an EM algorithm, based on the calculation of the forward and backward variables, $a_t(\mathbf{i}) = P(\mathbf{O_1}, ..., \mathbf{O_t}, \mathbf{q_t} = \mathbf{i})$ and $\beta_t(\mathbf{i}) = P(\mathbf{O_{t+1}}, ..., \mathbf{O_T}, \mathbf{q_t} = \mathbf{i})$ respectively, where T the length of the observation sequence:

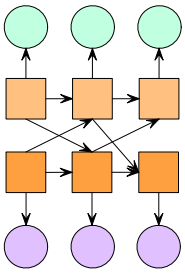


Fig. 19. Coupled Hidden Markov Model structure. Squares denote the hidden nodes of each interacting process and circles the associated observable outputs

$$a_1(\mathbf{i}) = \pi(\mathbf{i})b_1(\mathbf{i}) \tag{34}$$

$$a_t(\mathbf{i}) = b_t(\mathbf{i}) \sum_j \alpha(\mathbf{i}|\mathbf{j})a_{t-1}(\mathbf{j}) \tag{35}$$

for $t = 2, 3, ..., T$

$$\beta_1(\mathbf{i}) = 1 \tag{36}$$

$$\beta_t(\mathbf{j}) = \sum_i b_{t+1}(\mathbf{i})\alpha(\mathbf{i}|\mathbf{j})\beta_{t+1}(\mathbf{i}) \tag{37}$$

for $t = T, T-1, ..., 2$

The probability of the r_{th} observation sequence O_r of length T_r is computed as $a_{r,T}(N_1, N_2, ...N_S) = \beta_{r,1}(1, ..., 1)$

The scaled version of the forward and backward variables ($\hat{a}$,$\hat{\beta}$) [48] obtained in the E step are used to re-estimate the transition and observation parameters as follows:

$$\tilde{\alpha}^s(i|\mathbf{j}) = \frac{\sum_r \sum_{\mathbf{i},s.t.i_s=i} \sum_t \hat{a}_{r,t}(\mathbf{j})\alpha(\mathbf{i}|\mathbf{j})b_{r,t+1}(\mathbf{i})\hat{\beta}_{r,t+1}(\mathbf{i})}{\sum_r \sum_t \hat{a}_{r,t}(\mathbf{j})\hat{\beta}_{r,t}(\mathbf{j})\frac{1}{c_t}} \tag{38}$$

$$\tilde{b}^s_i(k) = \frac{\sum_r \sum_{\mathbf{i},s.t.i_s=i} \sum_{t,s.t.O^s_t=k} \hat{a}_t(\mathbf{i})\hat{\beta}_t(\mathbf{i})\frac{1}{c_t}}{\sum_r \sum_t \sum_{\mathbf{i},s.t.i_s=i} \hat{a}_t(\mathbf{i})\hat{\beta}_t(\mathbf{i})\frac{1}{c_t}} \tag{39}$$

where c_t the scaling coefficient for time t.

The number of states has been defined taking into consideration the inherent structure of each action. For instance, the Phone Conversation action consists of the "natural" states "Ringing" - "Reach Phone"- "Bring close to Head" - "Speech" - "Hang Up", upon which various second layer events can be defined.

C. Activity - Related Biometric Authentication Module

The aim of the activity - related biometric authentication module is to receive the dynamics of the human posture produced by the body motion tracking module on some user action segmented by the Activity Recognition Module and output some authentication results (Fig 18). Within this project we would like to evaluate the assumption that behavior can be employed as biometric signal as well as the hypothesis that our belief measure on the user's identity increases with time. Furthermore, various work - related motions should be tested with regard to their discriminative power.

Related work includes several model - based and feature - based methods for human gait identification and authentication [62], [13]. Key stroke dynamics have been also employed for activity - related person authentication [42]. To our knowledge, activity - related person authentication based on environment generated stimuli and work - related activities is a completely novel concept and has never been implemented before.

The output of this module for a particular action could either be a strict authentication result (Accepted/Rejected) or a belief measure that can be integrated with future partial inferences of the same modality and/or inferences of other modalities to converge in a final authentication result at later time stamps (Continuous Authentication). The latter approach seems more promising, as the user's "natural" behavior can be more reliably confirmed on multiple action instances. In general, a user's way

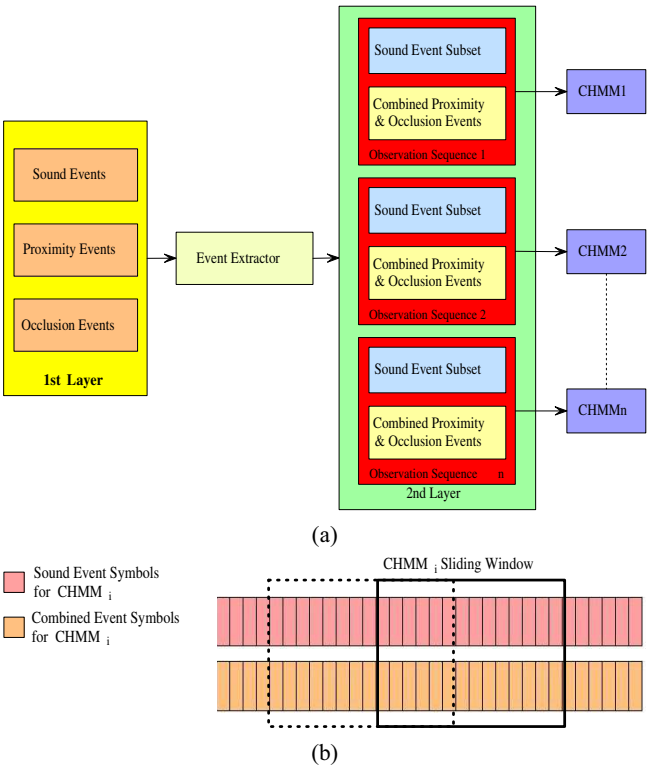


Fig. 20. a) Event Extraction b) Sliding Window for CHMM

of execution of some motion can diverge from its usual dynamics on single instances depending on various factors (psychological condition, unusual environmental conditions etc.). Despite that, it can be assumed that over longer periods of times where multiple instances of many actions take place, the user's identity could be reliably inferred.

The Activity - Related Biometric Authentication Module assumes a mapping of a user's behavior to his identity, therefore tools, methods and features that have been used for action and gesture recognition can be applied. In this implementation the body joint angles and position of the central point of the human model (III) and their derivatives are used as features for modeling the user's natural way of executing some action, since those features can powerfully represent the human model posture and its dynamics. Principal Component Analysis for each action class is used to reduce the dimensionality of the feature vector.

For biometric authentication Hidden Markov Models with Multivariate Gaussian outputs are used to capture the spatio - temporal dynamics of the human behavior. Standard HMM classification is performed by assigning one model to every individual enrolled in the authentication system. Given some extracted observation sequence $O_{1:T}$ of length T associated to a segmented action, and the set of HMMs $\lambda_i, i = 1, \ldots, N$ where N the number of enrolled users, the probability $P(O|\lambda_i)$ is calculated for all HMMs. By assigning an authentication threshold to each user's HMM, direct authentication results based on single actions can be obtained. A more promising option is propagating all the above probabilities to an integration module that emits authentication results on longer periods of activity. Fig. 21 graphically represents the Activity - Related Biometric Authentication Module.

X. Integration of Decisions

A typical authentication system presents a DET (Detection Error Trade-off) curve which enables a system to select a point on the curve to trade off between security and ease of use of a system. However, a continuous authentication system needs to traverse this DET curve based on the current situation. If the system is confident based on past inferences, temporary drops in the probability of the target class should not cause the rejection of the user. However if there is an elongated period of diffidence about the authenticity of the target person, then the system should be able to reject the person eventually.

The second problem is the integration of the inferences from the different modalities. Each mode produces different inferences with a

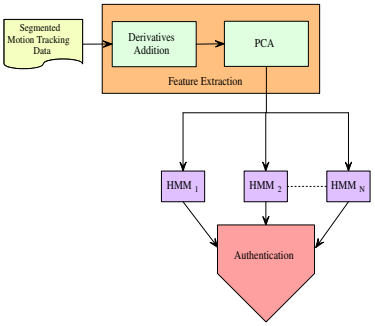


Fig. 21. Activity - related biometric feature extraction and authentication

different probability and these inferences are available at different points in time. There is the additional complication of assessing the reliability (and consequently the relative weight) of each modality. This problem is termed in the literature as "Holistic Fusion".

Among previous work done on holistic fusion, the most significant are Zhang *et al.* [67] and Kittler *et al.* [25]. Zhang *et al.* suggested a two state Hidden Markov Model, where the two states are "safe" and "attacked". A decay factor was proposed, which exponentially weighted over the previous observations, as well as weighted sums to integrate over modalities, where the weights were the assessed reliabilities of the modalities. The area under the Receiver Operating Characteristics (ROC) curve for each modality is used to quantify reliability. The approach we will present now is similar in some respects to this method, but it does not use HMMs.

Let λ_Ω be the model of 'target' person, the person whom we want to authenticate. Let λ_m be one among the M impostor models. Let O_t^n be the t^{th} observation in time, among Γ observations from the n^{th} modality among N modalities. Each module produces the likelihood of λ_Ω given O_t^n, i.e. $p(O_t^n|\lambda_\Omega)$. Since the likelihoods from different modalities have the inherent problem of being in different scales, it becomes difficult to find suitable weights. So the posterior is calculated as follows

$$P(\lambda_\Omega|O_t^n) = \frac{p(O_t^n|\lambda_\Omega) * P(\lambda_\Omega)}{p(O_t^n)} \tag{40}$$

The next question is about calculating the prior $P(\lambda_\Omega)$ and the observation probability $p(O_t^n)$. The observation probability can be given by

$$p(O_t^n) = p(O_t^n|\lambda_\Omega) * P(\lambda_\Omega) + \sum_{m=1}^{M} p(O_t^n|\lambda_m) * P(\lambda_m) \tag{41}$$

How to estimate $P(\lambda_\Omega)$ is an interesting problem. This value is tunable and different points on the DET curve can be achieved by changing this value. Increasing this value makes the system more confident about the authenticity of the subject and thereby increases the false acceptance rate (FAR). Reducing this value increases the false rejection rate (FRR) while decreasing the FAR.

A continuous authentication system is typically used after the authenticity is verified by an independent system. The initial estimate of the prior, $P_0(\lambda_\Omega)$, can be received from this entry system or taken to be an arbitrarily high value. The subsequent values of this prior are calculated as shown below

$$P_t(\lambda_\Omega) = \frac{\sum_{n=1}^{N} P_t(\lambda_\Omega|O_t^n) * P(O_{1:t}^n)}{\sum_{n=1}^{N} P(O_{1:t}^n)} \tag{42}$$

where

$$P(O_{1:t}^n) = \frac{\left(\frac{p(O_i^n|\lambda_\Omega) + \sum_{m=1}^{M} p(O_i^n|\lambda_m)}{M+1}\right)}{\frac{1}{t}\sum_{i=1}^{t}\left(\frac{p(O_i^n|\lambda_\Omega) + \sum_{m=1}^{M} p(O_i^n|\lambda_m)}{M+1}\right)} \tag{43}$$

Now with a time-varying estimate of prior available equation 40 can be combined with 41 and written as shown below.

$$P_t(\lambda_\Omega|O_t^n) = \frac{p(O_t^n|\lambda_\Omega) * P_{t-1}(\lambda_\Omega)}{p(O_t^n|\lambda_\Omega) * P_{t-1}(\lambda_\Omega) + \sum_{m=1}^{M} p(O_t^n|\lambda_m) * P_{t-1}(\lambda_m)} \tag{44}$$

where $\forall m$

$$P_{t-1}(\lambda_m) = \frac{1 - P_{t-1}(\lambda_\Omega)}{M} \tag{45}$$

The prior is updated at every calculation and the confidence of the system depends on all the different modalities. In a system such as the one described in the experiment, it may not be possible to get a new inference from each modality at each instance of time. So the latest inference from each modality is used for re-computing the estimate of the prior, λ_Ω.

The strategy proposed builds a confidence value about the identity of a person. This confidence is in terms of the updated posterior probability. If the different modalities ascribe low confidence to the authenticity of the person, then the overall confidence drops down. But if the modalities provide high confidence to the authenticity, then the overall confidence in the person builds up. At some point, if one of the modalities ascribes low confidence to the authenticity of the target, then it is weighed by how probable the occurrence of such an observation is. So if an observation is not very probable in the model of the entire system, then a lower weight is given in the overall confidence calculation.

If at any point, the user is switched with an impostor, it will take some time for the system to bring down the confidence levels due to the high confidence levels initially built on the user, and the impostor is likely to be authenticated for some time. But the overall confidence will drop eventually, with a speed that depends on the confidence scores of each modality. Using a window-approach that takes into account the last *k* frames in assessing probabilities may be useful in providing a fast decrease under switched persons.

Further testing needs to be done in the case of impostor switching and hysteresis of the system under these circumstances.

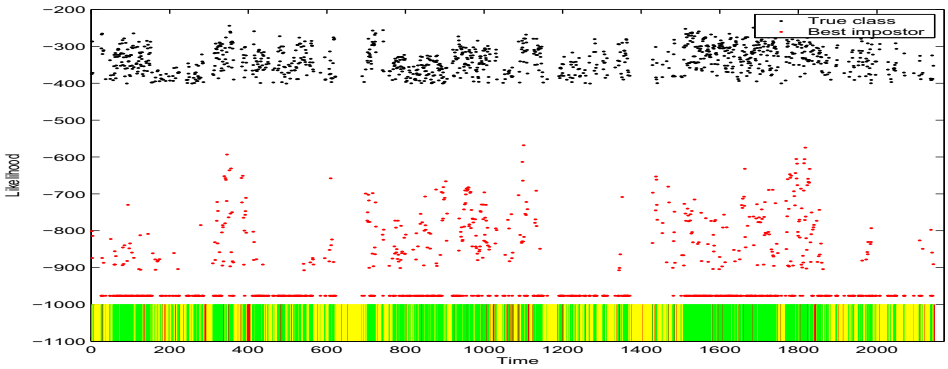
XI. Experimental Results

A. Continuous Face Authentication

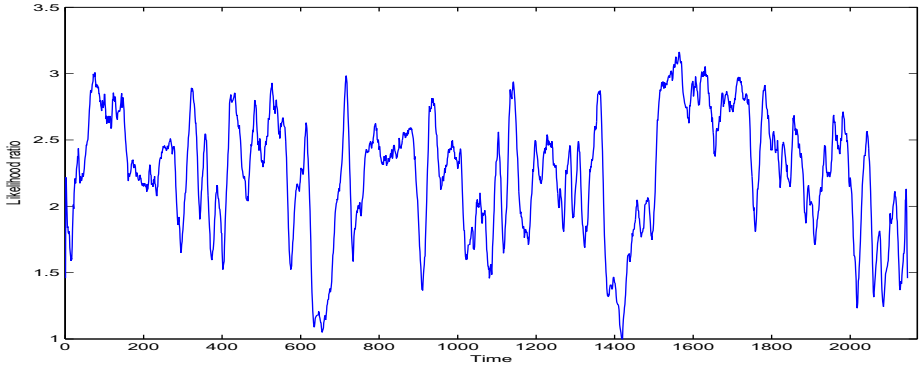
The face authentication module is tested with the recordings of 11 individuals. The first session is used to construct the statistical models for each person. The remaining nine sessions are used for reporting the success of the algorithm. For 99 sessions, the face detection module locates faces 92.3 per cent of the total recording time, with a standard deviation equal to 7.4 per cent. This means that for a 1000 frame session, about 923 face images are processed for authentication. Some of these faces are false alarms, caused by the failure of the Viola-Jones face detector.

In general, the face detection module is robust enough to correctly localize faces during activities like phone conversations. This implies that for these frames, the cropped face area contains the hand and the phone itself. We have observed that the face authentication module frequently stays below authentication threshold for these cases. Fig. 22 shows the authentication result for a single session. The horizontal axis is the time, and the vertical axis is the likelihood value obtained by the class models. Each face is shown as a dot on this plot. We only report the likelihood from the genuine class and the best impostor claim for that frame. The threshold is selected as -400, and the shown sequence justifies this choice nicely. In fact the threshold is optimized on a separate set, but since it strictly depends on the subspace dimensionality, it produces uniformly good results across the test sessions, as shown by the low variance of the results. At the bottom of the figure, a coloured band indicates when faces are not detected in the video (with red), when they are detected but the true class authentication does not follow (with yellow) and correct authentications (with green). The parts with longer bands of yellow are the activities where the face is not isolated or completely frontal.

The complete testing data consists of 91250 frames, recorded from nine sessions per subject, and 11 subjects. For each frame, the best impostor access is selected by evaluating the remaining 10 models. We demostrate the effect of selecting different thresholds in Fig. 23, where the false accept rate and the false rejection rate of the system are plotted for a range of threshold values. For the selected threshold of -400, the system has 0.3 per cent false acceptance rate and 30.1 false rejection rate. This means that for a video sequence with 1000 detected faces, roughly 3 frames would admit impostors, and 700 frames would indicate the presence of the true user. At this level, there is no interpretation of these results. In practice, a session of continuous authentication can



(a)



(b)

Fig. 22. The output of the continuous face authentication module during one session. (a) The likelihood of the genuine and best impostor claims. The band shows correct authentications (green), no authentication (yellow), and no detection (red) cases. (b) The likelihood ratio of the genuine class to the best impostor class for the same session.

operate on a sliding window of frames, where the genuine and impostor likelihoods are compared, and the system outputs a decision at every time slot. Under these controlled conditions (i.e. difficult but similar illumination conditions in training and test sessions), it is obvious that the face modality provides very robust authentication.

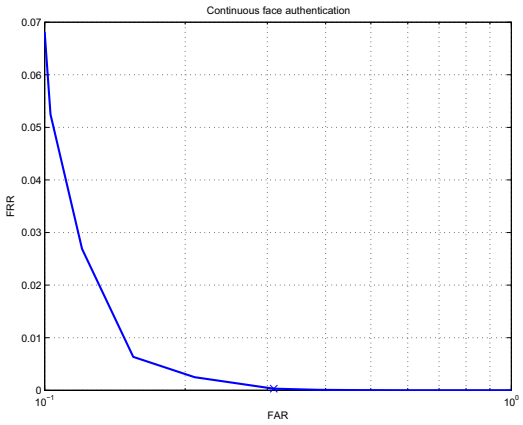


Fig. 23. The receiver operator characteristic curve for a range of thresholds of authentication. The genuine class is evaluated against the best impostor model for each frame. The average values for 99 sessions are reported. The cross indicates the selected threshold for the operating point of the system.

B. Speaker verification

For purposes of training and testing, approximately 20 seconds of speech is recorded during each session in form of a telephone conversation in addition to 40 seconds of speech in form of reading a paragraph

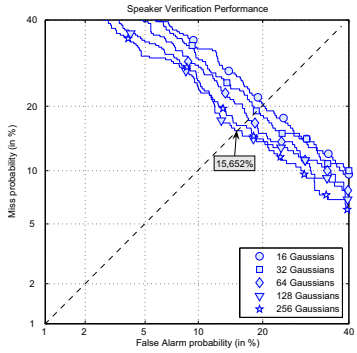


Fig. 24. DET curve for speaker verification module

of written text. 15 subjects have contributed to the database, with 10 recording sessions per subject. The results reported in this section are obtained by training with sessions one to five, and testing with the session six and seven, for 10 subjects. We have evaluated GMMs with different numbers of components.

Fig. 24 shows that the best results are achieved using 128 components for Gaussian mixture densities.

C. Contact-based Biometrics

The experimental setup includes one testing action and eight subjects. In particular, the right hand of each user and the glass of the office were registered in the virtual environment for the action notated as "grabbing the glass". For the classification we implemented standard techniques of pattern recognition. PCA was used to reduce the dimensionality of the feature space while neural networks were trained for the final classification. Each person performed the action 10 times which produced

1000 sample frames on average for each subject due to the high sampling frequency of the CyberGlove®. From these samples 70% were used to train the network and 30% for testing. Fig. 25 displays the final ROC curve of the FAR and FRR rates for the testing data of the eight subjects.

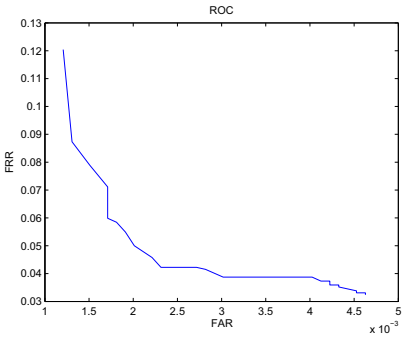


Fig. 25. ROC curve for the action "grabbing the glass" and eight subjects.

The results show that collision features are comparable to other Activity-Related biometrics and therefore comprise a very interesting approach for user authentication.

D. Body Motion Tracking

The body tracker was tested in the office pilot. Two webcams, one frontal and one lateral, recording at 9.5 fps provided the frames onto which the 3D articulated model was projected. 3D body part locations (head, shoulders, elbows and wrists) have been manually annotated in one subject sequence in order to test the tracker performance. The error is expressed as the mean distance between the annotated and the estimated joints. Comparative results between the APF with the common likelihood approach (comprising edges and foreground matching) and our proposal are shown in Fig. 26. In both cases we used the body model and the projection procedure explained in section III-C. Final mean error obtained by our approach for this sequence was 85 mm. Common likelihood evaluation makes the tracker vulnerable to track loss, leading to higher mean error. On the other hand, the divergence measure and the feature-based smoothing of the likelihood approximation make the tracker more robust under our experimental conditions.

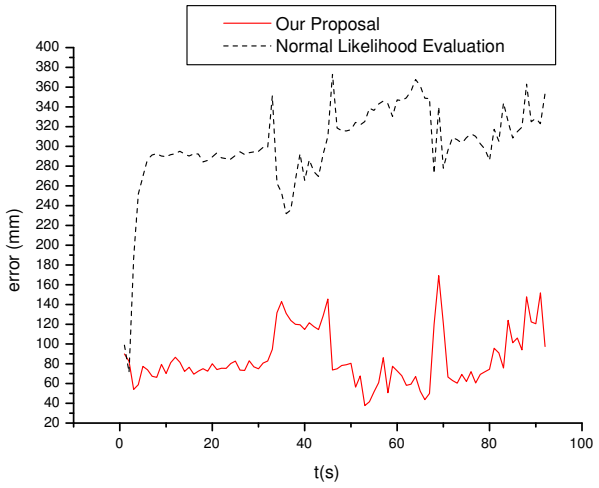


Fig. 26. Comparative results using 3 layers and 200 particles per layer with the normal likelihood aproximation and our proposal.

We found out that some spurious contours due to clothing and objects caused our tracker to fail in its estimation. The aparent motion recorded in the images was very fast in some of the actions required for activity-based recognition. These aparent fast motions caused blurs in the image and abrupt translation of body parts. Since the implemented annealing PF works with contours as most determinant feature, the algorithm was not able to track several of these fast motions. However, it was able to recover some poses after a tracking error. Similarly, we detected that some poses couldn't be retrieved due to self-occlusions caused by the lack of additional views. Therefore, for some of the actions and poses, the problem becomes ill-posed and, as a consequence, more information is needed.

After testing several sequences, it was found that for several non-fast motions good results can be obtained with 3 layers and between 100 and 200 particles per layer. However, a more exhaustive study with ground truth angles must be done under similar conditions in order to refine the likelihood approximation, the annealing parameters and the number of particles.

E. Other Modules

The results of the Sound - based Event Detection module are illustrated in the respective section IV. Testing of the Activity Recognition module and the Behavioral Face Biometrics module remain as future work.

Preliminary results for the Activity - related biometric module reveal a potential of using work - related activities as biometric signals. Experimenting on 7 manually segmented sequences (5 for training and 2 for testing) of the action classes Writing and Phone Conversation, we found out that the true person receives good HMM log - likelihood ranking. Despite that, the need for more accurate and stable 3D Motion Tracking was obvious, as it is the case for most state of the art model - based techniques. Future work includes testing on larger sets and more action classes, with improved motion tracking data. A feature - based approach (direct feature extraction on segmented human blobs) will also be implemented.

XII. Conclusions and Future Directions

In this project we have evaluated several activity related biometric modalities for their relative success in continuously determining and verifying the identity of a user in a typical and non-obtrusive work environment scenario. Apart from more traditional face and speech based verification, facial actions and movement patterns were assessed for authentication. A pilot setup with different action scenarios is defined, and a large database is collected from 15 subjects. Each subject contributed 10 sessions, which are manually annotated by the project group for further evaluation.

The experimental evaluation of all the modalities is not achieved exhaustively, and their possible integration remains to be a future endeavor. The latter is partly due to the success of individual modalities on the restricted pilot setup, which suggests that under closely resembling training and testing conditions there will be no marked benefit under fusion scenarios. However, the results demonstrate that activity-based biometrics is a promising venue for further study.

XIII. Acknowledgements

The authors thank Christophe D'Alessandro and the organization team of the eNTERFACE'08 for all their efforts. Albert Ali Salah is supported by the Dutch BRICKS/BSIK project, and a scientific mission grant from EU COST 2101 Action. Martin Lojka is supported by Ministry of education of Slovak Republic under research project VEGA 1/4054/07 and Slovak Research and Development Agency under research project APVV-0369-07. This work was supported in part by the EC under contract FP7-215372 ACTIBIO.

References

[1] G. Aggarwal, A. Chowdhury, and R. Chellappa. A system identification approach for video-based face recognition. *Proc. Int. Conf. on Pattern Recognition*, 4:176–178, 2004.

[2] M. Arulampalam, S. Maskell, N. Gordon, T. Clapp, D. Sci, T. Organ, and S. Adelaide. A tutorial on particle filters for online nonlinear/non-GaussianBayesian tracking. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 50(2):174–188, 2002.

[3] E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, et al. The BANCA Database and Evaluation Protocol. *LNCS*, pages 625–638, 2003.

[4] P. Balucha. Automatic speaker recognition with gaussian mixture models. Master's thesis, Technical University of Košice, 2006.

[5] F. Bashir, A. Khokhar, and D. Schonfeld. Object trajectory-based activity classification and recognition using hidden markov models. *Image Processing, IEEE Transactions on*, 16(7):1912–1919, July 2007.

[6] N. Boulgouris and Z. Chi. Gait Recognition Using Radon Transform and Linear Discriminant Analysis. *IEEE Transactions ON Image Processing*, 16(3):731, 2007.

[7] C. Bregler and J. Malik. Tracking People with Twists and Exponential Maps. *IEEE Computer Society Conference On Computer Vision And Pattern Recognition*, 1998.

[8] A. Bregman. *Auditory scene analysis*. MIT Press Cambridge, Mass, 1990.

[9] F. Caillette, A. Galata, and T. Howard. Real-Time 3-D Human Body Tracking using Variable Length Markov Models. *British Machine Vision Conference*, 1:469–478, 2005.

[10] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[11] C. Canton-Ferrer, J. Casas, and M. Pardas. Exploiting Structural Hierarchy in Articulated Objects Towards Robust Motion Capture. *Lecture Notes in Computer Science*, pages 82–91, 2008.

[12] L. Chen, H. Liao, and J. Lin. Person identification using facial motion. *Image Processing, 2001. Proceedings. 2001 International Conference on*, 2, 2001.

[13] M.-H. Cheng, M.-F. Ho, and C.-L. Huang. Gait analysis for human identification through manifold learning and hmm. *Pattern Recogn.*, 41(8):2541–2553, 2008.

[14] A. Colmenarez, B. Frey, and T. Huang. A Probabilistic Framework for Embedded Face and Facial Expression Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:592–597, 1999.

[15] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. *PROC IEEE Comput Soc Conf Comput Vision Pattern Recognit*, 2:126–133, 2000.

[16] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, July 2000.

[17] C. Ericson. *Real-Time Collision Detection*. Morgan Kaufmann, 2004.

[18] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113, 1993.

[19] K. Huang and M. Trivedi. Streaming face recognition using multicamera video arrays. *Proc. Int. Conf. on Pattern Recognition*, 4:213–216, 2002.

[20] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing*. Prentice Hall, 2001.

[21] M. Isard and A. Blake. CONDENSATION-Conditional density propagation for visual tracking. *Int. Journal of Computer Vision*, 29(1):5–28, 1998.

[22] A. Jain, S. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. *Lecture notes in computer science*, pages 731–738.

[23] A. Kale, N. Cuntoor, and R. Chellappa. A framework for activity-specific human recognition. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (Orlando, FL)*, 706, 2002.

[24] Y. J. Kim, M. C. Lin, and D. Manocha. Incremental penetration depth estimation between convex polytopes using dual-space expansion. *IEEE Transactions on Visualization and Computer Graphics*, 10(2):152–163, 2004.

[25] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:226–239, 1998.

[26] E. Larsen, S. Gottschalk, M. Lin, and D. Manocha. Fast distance queries with rectangular swept sphere volumes. volume 4, pages 3719–3726, 2000.

[27] A. Laurentini. The visual hull concept for silhouette-based image understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(2):150–162, 1994.

[28] K. Lee, J. Ho, M. Yang, and D. Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, 99:303–331, 2005.

[29] B. Li and R. Chellappa. A generic approach to simultaneous tracking and verification in video. *IEEE Transactions on Image Processing*, 11:530–544, 2002.

[30] B. Li, R. Chellappa, Q. Zheng, and S. Der. Model-based temporal object verification using video. *IEEE Transactions on Image Processing*, 10(6):897–908, 2001.

[31] Z. Li, S. Wachsmuth, J. Fritsch, and G. Sagerer. View-adaptive manipulative action recognition for robot companions. *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, pages 1028–1033, 29 2007-Nov. 2 2007.

[32] M. C. Lin and J. F. Canny. A fast algorithm for incremental distance calculation. pages 1008–1014, 1991.

[33] M. C. Lin and S. Gottschalk. Collision detection between geometric models: A survey. In *In Proc. of IMA Conference on Mathematics of Surfaces*, pages 37–56, 1998.

[34] J. Liu and R. Chen. Sequential Monte Carlo methods for dynamical systems. *Journal of the American Statistical Association*, 93(5):1032–1044, 1998.

[35] X. Liu and T. Chen. Video-Based Face Recognition Using Adaptive Hidden Markov Models. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1, 2003.

[36] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *DARPA Image Understanding Workshop*, pages 121–130, 1981.

[37] J. MacCormick and M. Isard. Partitioned Sampling, Articulated Objects, and Interface-Quality Hand Tracking. *Lecture Notes in Computer Science*, pages 3–19, 2000.

[38] B. Mak, J.-C. Junqua, and B. Reaves. A robust speech/non-speech detection algorithm using time and frequency-based features. *Proc. ICASSP*, 1:269–272, 1992.

[39] F. Matta and J. Dugelay. A behavioural approach to person recognition. *Proceedings of IEEE International Conference on Multimedia and Expo (ICME2006)*, pages 9–12, 2006.

[40] I. Mikic. Human Body Model Acquisition and tracking using multi-camera voxel Data. *PhD. Thesis, University of California, San Diego*, 2003.

[41] J. Mitchelson and A. Hilton. Simultaneous pose estimation of multiple people using multiple-view cues with hierarchical sampling. *In Proc. of BMVC, September*, 2003.

[42] F. Monrose and A. Rubin. Authentication via keystroke dynamics. In *CCS '97: Proceedings of the 4th ACM conference on Computer and communications security*, pages 48–56, New York, NY, USA, 1997. ACM.

[43] D. J. Moorey, I. A. Essaz, and M. H. H. Iiiy. Objectspaces: Context management for human activity recognition. In *Second International Conference on Audio- Vision-based Person Authentication*, 1999.

[44] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP J. Appl. Signal Process.*, 2002(1):1274–1288, 2002.

[45] E. Nemer, R. Goubran, and S. Mahmoud. Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Transactions on Speech and Audio Processing*, 9:217–231, 2001.

[46] D. Patterson, D. Fox, H. Kautz, and M. Philipose. Fine-grained activity recognition by aggregating abstract object usage. *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on*, pages 44–51, Oct. 2005.

[47] P. Psutka, L. Müller, J. Matoušek, and V. Radová. *Mluvíme s počítačem česky*. Academia, 2004.

[48] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. pages 267–296, 1990.

[49] N. Ramanathan, R. Chellappa, and A. Roy Chowdhury. Facial similarity across age, disguise, illumination and pose. *Proc. Int. Conf. on Image Processing*, 3, 2004.

[50] L. Raskin, E. Rivlin, and M. Rudzsky. Using Gaussian Process Annealing Particle Filter for 3D Human Tracking-Volume 2008, Article ID 592081, 13 pages. *EURASIP Journal on Advances in Signal Processing*, 2008.

[51] A. Raynolds. Speaker identification and verification using Gaussian mixture speaker models.

[52] P. Renevey and A. Drygajlo. Entropy Based Voice Activity Detection in Very Noisy Conditions. In *Seventh European Conference on Speech Communication and Technology*. ISCA, 2001.

[53] A. Salah and E. Alpaydın. Incremental mixtures of factor analyzers. *Int. Conf. on Pattern Recognition*, 1:276–279, 2004.

[54] M. Savvides and B. Vijaya Kumar. Illumination normalization using Logarithm transforms for face authentication. *Lecture notes in computer science*, pages 549–556.

[55] M. Stager, P. Lukowicz, and G. Troster. Implementation and evaluation of a low-power sound-based user activity recognition system. In *ISWC '04: Proceedings of the Eighth International*

Symposium on Wearable Computers, pages 138–141, Washington, DC, USA, 2004. IEEE Computer Society.

[56] C. Stauffer and W. Grimson. Learning Patterns of Activity Using Real-Time Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 747–757, 2000.

[57] M. Teschner, S. Kimmerle, G. Zachmann, B. Heidelberger, L. Raghupathi, A. Fuhrmann, M.-P. Cani, F. Faure, N. Magnetat-Thalmann, and W. Strasser. Collision detection for deformable objects. In *Eurographics State-of-the-Art Report (EG-STAR)*, pages 119–139. Eurographics Association, 2004.

[58] F. Thomas and C. Torras. 3d collision detection: A survey. *Computers and Graphics*, 25:269–285, 2001.

[59] G. Van and D. Bergen. Efficient collision detection of complex deformable models using aabb trees. *J. Graphics Tools*, 2, 1997.

[60] G. van den Bergen. *Collision Detection in Interactive 3D Environments*. Morgan Kaufmann, 2003.

[61] P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. *IEEE Computer Society Conference On Computer Vision And Pattern Recognition*, 1, 2001.

[62] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, 2003.

[63] J. A. Ward, P. Lukowicz, G. Troster, and T. E. Starner. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1553–1567, 2006.

[64] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg. A scalable approach to activity recognition based on object use. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct. 2007.

[65] L. Xu, J. Landabaso, and M. Pardas. Shadow Removal with Blob-Based Morphological Reconstruction for Error Correction. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2, 2005.

[66] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge University Engineering Department, 2006.

[67] S. Zhang, R. Janakiraman, T. Sim, and S. Kumar. Continuous Verification Using Multimodal Biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:687–700, 2007.

Gopal Ananthakrishnan received his Master of Science (Engg) from the Indian Institute of Science, Bangalore, India, in 2007. Currently, he is a PhD candidate in the Royal Institute of Technology, Stockholm, Sweden. His main research interests include Auditory to Articulatory Inversion, Speech production, Analysis of Speech and Audio signals, Perceptual and Neuro-Physiological analysis of hearing, Modeling and Simulation of Perceptual properties of human hearing, Pattern Recognition, On-line Handwriting Recognition.

Hamdi Dibeklioğlu was born in Denizli, Turkey in 1983. He received his B.Sc. degree from Yeditepe University Computer Engineering Department, in June 2006, and his M.Sc. degree from Boğaziçi University Computer Engineering Department, in July 2008. He is currently a research assistant and a Ph.D. student at Boğaziçi University Computer Engineering Department. His research interests include 3D face recognition, computer vision, pattern recognition and intelligent human-computer interfaces. He works with Professor Lale Akarun on 3D Face Recognition.

Martin Lojka was born in Snina, Slovakia, in 1984. He received engineer's degree in 2007. He is currently a PhD student in department of Electronics & Multimedia Communications of Technical University of Košice under the supervision of Doc. Ing. Jozef Juhar, PhD. His research interests focus on algorithms of audio processing in embedded systems.

Adolfo López was in born in Barcelona in 1982. He received his BS in Telecommunication Engineering from Universitat Politècnica de Catalunya (UPC) in Barcelona in 2007. He is currently a PhD student in the Image and Video Processing Group of the UPC under the supervision of professor Josep Ramon Casas. His research interests include Markerless Body Pose and Motion Estimation, Gesture and Motion Recognition and Particle Filter based Visual Tracking.

Serafeim Perdikis received his Diploma in Electrical and Computer Engineering from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2007. Currently, he is a PhD candidate in the Aristotle University of Thessaloniki and a research fellow with the Informatics and Telematics Institute, Centre for Research and Technology Hellas, Thessaloniki. His main research interests include activity recognition, gesture recognition, human - computer interaction, dynamic biometrics and signal processing. He is also a member of the Technical Chamber of Greece.

Usman Saaed was born in Lahore, Pakistan in 1981. He received a BS in Computer System Engineering from GIK Institute (Topi, Pakistan) in 2004. After graduation he was associated with the Electrical Engineering Dept. of Comsats Institute (Lahore, Pakistan) as a research associate. In 2005, he joined the University of Nice-Sophia Antipolis (Sophia Antipolis, France) for a Master of Research in Image Processing. He is currently a PhD student in the Multimedia Communication department of Institut Eurecom (Sophia Antipolis, France) under the supervision of Prof. Jean- Luc Dugelay. His current research interests focus on facial analysis in video.

Albert Ali Salah received his PhD in 2007 from the Dept. of Computer Engineering of Boğaziçi University, with a dissertation on biologically inspired 3D face recognition. This work was supported by two FP6 networks of excellence: BIOSECURE on multimodal biometrics, and SIMILAR on human-computer interaction. His research areas are pattern recognition, biometrics, and multimodal information processing. He received the inaugural EBF Biometrics Research Award in 2006, and joined with the Signals and Images group at CWI, Amsterdam as a BRICKS scholar. Recent scientific assignments include program committee memberships for BIOID'08, biometrics track of ICPR'08, and ICB'09.

Dr. Dimitrios Tzovaras is a Senior Researcher (Grade B) at the Informatics and Telematics Institute. He received the Diploma in Electrical Engineering and the Ph.D. in 2D and 3D Image Compression from the Aristotle University of Thessaloniki, Greece in 1992 and 1997, respectively. Prior to his current position, he was a senior researcher on the Information Processing Laboratory at the Electrical and Computer Engineering Department of the Aristotle University of Thessaloniki. His main research interests include information management, multimodal data fusion and knowledge management. His involvement with those research areas has led to the coauthoring of over thirty articles in refereed journals and more than eighty papers in international conferences. He has served as a regular reviewer for a number of international journals and conferences. Dr. Tzovaras is an Associate Editor of the Journal of Applied Signal Processing (JASP).

Athanasios Vogiannou received his Diploma degree in electrical and computer engineering from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2007. Currently, he is a PhD candidate in the Aristotle University of Thessaloniki and a research fellow with the Informatics and Telematics Institute, Centre for Research and Technology Hellas, Thessaloniki. His main research interests include virtual reality, collision detection, physically based modeling, real time simulations and computer vision. He is also a member of the Technical Chamber of Greece.