Computer Vision and Image Understanding 116 (2012) 411–421

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

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

Spatiotemporal analysis of human activities for biometric authentication

Anastasios Drosou [a,b,*], Dimosthenis Ioannidis [b], Konstantinos Moustakas [b], Dimitrios Tzovaras [b]

[a] Imperial College London, SW7 2AZ London, UK
[b] Centre for Research & Technology Hellas (Ce.R.T.H.), Informatics & Telematics Institute (I.T.I), P.O. Box 60361, 57001 Thermi-Thessaloniki, Greece

ARTICLE INFO

Article history:
Available online 24 October 2011

Keywords:
Activity detection
Biometrics
Behavioral biometrics
Activity related authentication
HMM
Anthropometric profile
Attributed graph matching
Motion analysis
Body tracking

ABSTRACT

This paper presents a novel framework for unobtrusive biometric authentication based on the spatiotemporal analysis of human activities. Initially, the subject's actions that are recorded by a stereoscopic camera, are detected utilizing motion history images. Then, two novel unobtrusive biometric traits are proposed, namely the static anthropometric profile that accurately encodes the inter-subject variability with respect to human body dimensions, while the activity related trait that is based on dynamic motion trajectories encodes the behavioral inter-subject variability for performing a specific action. Subsequently, score level fusion is performed via support vector machines. Finally, an ergonomics-based quality indicator is introduced for the evaluation of the authentication potential for a specific trial. Experimental validation on data from two different datasets, illustrates the significant biometric authentication potential of the proposed framework in realistic scenarios, whereby the user is unobtrusively observed, while the use of the static anthropometric profile is seen to significantly improve performance with respect to state-of-the-art approaches.

1. Introduction

Human recognition has always been a field of primary concern in applications such as access control in secure infrastructures. In this respect, biometrics have recently gained significant attention from researchers, while they have been rapidly developed for various commercial applications, ranging from surveillance and access control against potential impostors to the management of voters to ensure no one votes twice [1,2]. Reliable personal recognition schemes are thus required to either confirm or determine the identity of an individual requesting their services. In the following, the linking between activity detection and behavioral based recognition of humans is attempted via a short introduction, since both scientific fields are combined in the framework presented herein.

1.1. Activity detection

Activity detection can be either performed as a pre-processing step in behavioral analysis or as a stand-alone application for triggering several types of alarm in surveillance areas (i.e. detecting suspicious movements or eventual threats and taking up the proper actions to prevent unwanted effects). In both cases, the main challenge lies in the segmentation of the activity of interest in a given frame sequence [3]. In other words, it is in some sense necessary to identify the portion of the sequence supposedly corresponding to an action, in order to recognize each action that appears in a given sequence.

In this respect, monitoring of activities has been extensively researched during the last decades, either for identifying ongoing events and activities [4], or for detecting anomalies in their execution [5], or for providing access control to restricted areas, etc. The results of this research field can be directly applied to surveillance [6], identity verification [7] and in medical applications [8].

In general, activity recognition methods can be divided in two main categories: (i) sensor-based methods and (ii) video-based methods. Regarding the first category, the authors of [9] mounted multiple accelerometers and a small low-power sensor board on single or multiple locations on the subject's body in order to estimate activities such as standing, walking or running. In the same concept, a hybrid multimodal approach for the automatic monitoring of everyday activities of elderly people was suggested in [4], whereby video analysis from multiple cameras was installed in an apartment and combined with information from sensors installed on doors, windows and the furniture. Similarly, a system for recognizing activities in the home setting using a set of state-change sensors was introduced in [10].

On the other hand, a real-time, less obtrusive, video understanding system was presented in [11], which automatically recognizes activities occurring in environments observed through video surveillance cameras. In the same respect, Wang et al. [12] managed to cluster spatiotemporal action of low dimensionality into manifolds using Locality Projective Projections (LPP). Thus, the detection of outdoor activities was achieved via geometrical simi-

* Corresponding author at: Centre for Research & Technology Hellas (Ce.R.T.H.), Informatics & Telematics Institute (I.T.I), P.O. Box 60361, 57001 Thermi-Thessaloniki, Greece.
E-mail address: drosou@iti.gr (A. Drosou).

1077-3142/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2011.08.009

larity measures between manifolds. Going a step further, Junejo et al. presented in [13] a novel vision based method for view-invariant action recognition by exploiting the advantages of temporal self-similarity matrices (SSM), as they are derived from the joints' extracted motion trajectories.

1.2. Biometric authentication

The variations exhibited by different users in the execution of the same activity [14] forms the main difficulty for an activity detection system. However, this significant issue for activity recognition systems turns into a fundamental advantage when dealing with user recognition systems, that try to exploit these variations to identify/authenticate individuals [14].

In this concept, a number of approaches have been described in the past to satisfy the different requirements of each application such as reliability, unobtrusiveness, permanence, etc. In general, biometric methods are categorized to (a) static or physiological and (b) behavioral human biometric characteristics [15], depending on the type of features used.

Physiological biometrics are based on biological measurements and inherent characteristics of each human. Static biometrics include fingerprint [16], DNA, face [17], iris and or retina [18] and hand geometry [19] or palm print [20] recognition. Despite their high accuracy, a general shortcoming of these biometric traits is their obtrusive process of obtaining the biometric feature. The subject has to stop, go through a specific measurement procedure, which can be very uncomfortable, wait for a period of time and get clearance after authentication is positive. Besides being obtrusive and uncomfortable for the user, static physical characteristics can be digitally duplicated, i.e. the face could be copied using a photograph, a voice print using a voice recording and the fingerprint using various forging methods. Moreover, static biometrics could be intolerant of changes in physiology such as daily voice changes or appearance changes.

On the other hand, behavioral biometrics, could overcome these drawbacks since they are related to specific actions and the way that each person executes them [21]. They can potentially allow the non-stop (on-the-move) authentication or even identification in an unobtrusive and transparent manner to the subject and become part of an ambient intelligence environment. Behavioral biometrics are the newest technology in biometrics and they have yet to be researched in detail. Even if physiological biometrics are considered more reliable, behavioral biometrics have the inherent advantage of being less obtrusive [22] and simpler to implement [15].

Similarly to the categorization of action recognition methods, previous work on human identification using behavioral signals can be roughly divided into: (a) sensor-based recognition [23] and (b) vision-based recognition. Recent research trends have been mainly moving towards the second category, due to various reasons, such as increased unobtrusiveness [24], continuous authentication capability, etc. Additionally, recent work and efforts on human recognition have shown that the human behavior (i.e. extraction of facial dynamics features [25]) and motion exploiting human body shape dynamics during gait ([26] or joints tracking analysis [27]), provide the potential of continuous authentication for discriminating people, when considering behavioral signals.

Gait, the first and most famous behavioral biometric trait since the late 1990s [28], has advantage of being a frequent periodic activity. In the same concept, recent works and efforts on human recognition have shown that there is high discriminative capacity in a series of other regularly executed activities [29] that also exhibit significant potential towards unobtrusive, continuous user authentication.

As the reader would notice, vision-based biometric methods require high precision tracking algorithms in order to accurately capture the humans movements. In this respect, several tracking methods have been proposed in the bibliography, mainly dealing with gestural analysis of the human body [30]. Some exemplary works are shortly described hereafter.

An early and one of the most promising approaches towards initializing the upper-body shape has been proposed by Plankers and Fua [31], whereby an implicit ellipsoidal metaball representation has been fitted to stereo point clouds prior to tracking. However, by using a single pose which is updated at each time step there lies always the danger of tracking failure with a rapid movement or visual ambiguities pose estimation. Later, Ziegler et al. managed to map each point of an articulated human model on the corresponding 3*D* point cloud, derived from disparity data, for long sequences [32]. However the slow performance ($\sim$1 fps) that has been achieved, made this method unappropriate for real-time applications. A great improvement has been performed by Micilotta et al. [33], whereby the 3*D* gesture could be estimated in real time from 2*D* color images, by utilizing adaBoost trained detectors combined with heuristic and hard anthropometric rules. Still, the fact that they use color images makes it vulnerable to bad illumination or other lightning problems. Last but not least, a novel method for view invariant gesture recognition, utilizing spherical harmonics on high accuracy depth data (ToF camera) has been presented in [34].

1.3. Motivation

Recent trends in biometrics research deal with the analysis of the dynamic nature of various modalities targeting at users' convenience and optimal performance in various realistic environments. The idea behind using activity-related biometrics for recognition purposes is based on the observation that complex multijoint movements, such as walking or reaching an object, are planned and executed according to one's personal behavior and style. Furthermore, a number of natural ''restrictions'', such as the physiology of the human body, possible impairments or the perceived environment [35] are bound to influence constantly the type and the art of specific movements. Thus, it can be claimed that biometric recognition would be potentially feasible by basing on all these dynamic environmentally invariant properties (i.e. movement's distance, direction, starting/ending position, external load, etc.) [36].

One of the initial approaches to activity-related biometrics has been attempted in [21]. Specifically, Kale et al. measured signals from various modalities, while the subject performed various activities during walking. Then, these signals have been used to create either unimodal or multimodal activity-related biometric signatures for each subject. In this concept, the potential towards robust discrimination between subjects, as well as the persistence in some movements' characteristics has been exhibited.

Based on the above concept, a novel method for activity-related biometric authentication is proposed and demonstrated in the context of an office environment, within the current paper. In particular, the users are authenticated by analyzing the invariant features of their movement during several office related activities (i.e. a phone conversation and an interaction with a microphone panel). The analysis of the movements is based on the processing of the extracted motion trajectories, in order to retrieve unique signatures of dynamic nature that would form reliable biometric traits for authentication. The authentication performance of these dynamic traits is further augmented by exploiting the static anthropometric profile of the users' upper-body.

It can be claimed that the *Universality* requirement is by definition satisfied within the current approach, provided that a valid biometric characteristic should satisfy the following set of requirements [15]:

- Universality: Each user should posses it.
- Distinctiveness: The extracted features are characterized by great inter-individual differences.
- Reproducibility: The extracted features are characterized by small intra-individual differences.
- Permanence: No significantly change over time, age, environment conditions or other variables.
- Collectability & Automatic processing: Recognize or verify a human characteristic, which can be measured quantitatively, in a reasonable time and without a high level of human involvement.
- Circumvention: It should be difficult to alter or reproduce by an impostor who wants to fool the system.

Moreover, there is a plenty of models that indicate the users' tendency of seeking the "most convenient" or the least effort-demanding pose when performing a movement. In particular, according to *Flash and Hogan's Minimum Jerk Model* [37], the hand paths in extrinsic space should be straight, while curved hand paths can be generated as a concatenation of straight-line segments. Similarly, the *Uno, Kawato and Suzuki Minimum Torque Change Model* [38] assumed a hand movement according to the minimization of the torque during the movement. Based on these observations, but also on Turvey et al.'s [35] and Goodman et al.'s [36] findings, it can be proved that the *Distinctiveness*, *Reproducibility* but also the *Permanence* requirements are fulfilled as well, since all these parameters are related to the user's anthropometric variables, that exhibit significant variance within the population. Moreover, *Permanence* is guaranteed, given that an adults' body remains rather unchanged over the years, in terms of the distances between the joints. Finally, the combination of physiological with stylish and behavioral characteristics is bound to be very *hard to circumvent* or to mimic by an impostor.

A few further issues that should be considered when designing a practical biometric system are its recognition performance (i.e. accuracy, speed), the resources required to achieve the desired recognition accuracy and speed, as well as some operational and environmental factors (i.e. frequency of the performed activity on daily basis and the degree of societal approval). Given that the proposed method is totally unobtrusive to the user, but also given that the recognition process is incorporated in the users' everyday activities, it can be stated that the acceptability and frequency criteria are covered to an accepted extent. Last, the *Automatic processing* requirements including the recognition accuracy and the speed are highly dependent on the features and algorithms deployed and will be presented in the following.

In this respect, the current paper extends the activity-related biometric framework proposed in [39] by utilizing a new, faster and more accurate method for the extraction of the static anthropometric profile. A quantitative comparison between the two methods in terms of speed and recognition capacity is included. Compared to [39] the proposed framework is herein evaluated also with respect to a further office activity, the interaction (i.e. approaching and talking) with a microphone panel.

The rest of the paper is organized as follows: In Section 2, the overview of the proposed system is presented. The architecture proposed for the event detection framework is described in Section 3.1, while in Section 3.2 the basis for the current authentication approach is briefly reviewed. The contribution of static biometric information to an existing dynamic biometric system is presented in Section 3.3, followed by the introduction of a quality factor estimation based on the human ergonomics in Section 3.5. Finally, a short description of the database used and the experimental validation of our framework are presented and thoroughly discussed in Section 4.2.

2. Framework overview

An overview of the proposed biometric system is depicted in Fig. 1. Initially, the event detection module identifies and extracts by annotating the image sequence that corresponds to a specific activity. This sequence is then processed, so as to extract the user's static anthropometric profile that refers to the upper-body's skeleton model and the activity related dynamic features. The latter refer to information provided by the motion trajectories of the head and the hands. The full signature for the claimed user's ID is restored from the database and the actual extracted features are used to classify the user as a client or an impostor to the system via a Hidden Markov Model (HMM) classifier, concerning the dynamic motion, and an Attributed Graph Matcher (AGM), concerning the static anthropometric information. Finally, a score level fusion of both classifiers is performed by a Support Vector Machine (SVM) algorithm implemented on a Gaussian kernel and the validity of the final score is then verified by a quality factor based on ergonomic restrictions.

3. Activity related biometric authentication

In the following paragraphs follows an extensive description of each of the modules that form the proposed framework. The *Activity Recognition* module acts also auxiliary to the *Biometric Recognition* module, aiming primarily at verifying the integrity of the ongoing movement. Thus, if the phone falls from the user's hands, during a phone conversation, an alarm event will be raised and thus the current activity will be considered as non-valid for biometric recognition purposes.

3.1. Activity detection

Activity recognition is performed utilizing the concept of Motion History Images (MHI) [40]. Specifically, a MHI is a temporal template, where the intensity value at each point is a function of the motion properties at the corresponding spatial location in an image sequence according to Eq. (1)

$$MHI_T(x,y,t) = \begin{cases} \tau, & \text{if } D(x,y,t) = 1 \\ max(0, MHI_T(x,y,t-1) - 1), & \text{otherwise} \end{cases} \quad (1)$$

where τ is the number of frames contributing to the MHI generation and is proportional to the duration of the detected event (i.e. $\tau = 15$ for a *panel interaction* activity and $\tau = 30$ for a *phone conversation* activity). Further, $D(x,y,t)$ equals 1 if there is a difference in the intensity of a pixel between two successive frames $I_{t-1}(x,y)$; $I_t(x,y)$. The older a change is, the darker its depiction on the MHI will be, while changes older than 15 frames are discarded (Fig. 2).

The proposed system for activity recognition is presented in Fig. 3.

A MHI is extracted for each frame, using the last τ frames. The value of τ is different for each activity and has been experimentally selected. The criterion used, was the required maximum time needed for an activity to be completed by the user's. Given the camera's frame rate the value of τ was calculated. The general rule is indicates that the longer the duration of an activity, the bigger the value of τ.

Then, the MHI is transformed according to the Radial Integration Transform (RIT) and the Circular Integration Transform (CIT), which are used due to their aptitude to represent meaningful shape characteristics. The location of the head (x_0, y_0) is detected following the approach in [41] and is used as the center of integration for both transformations.

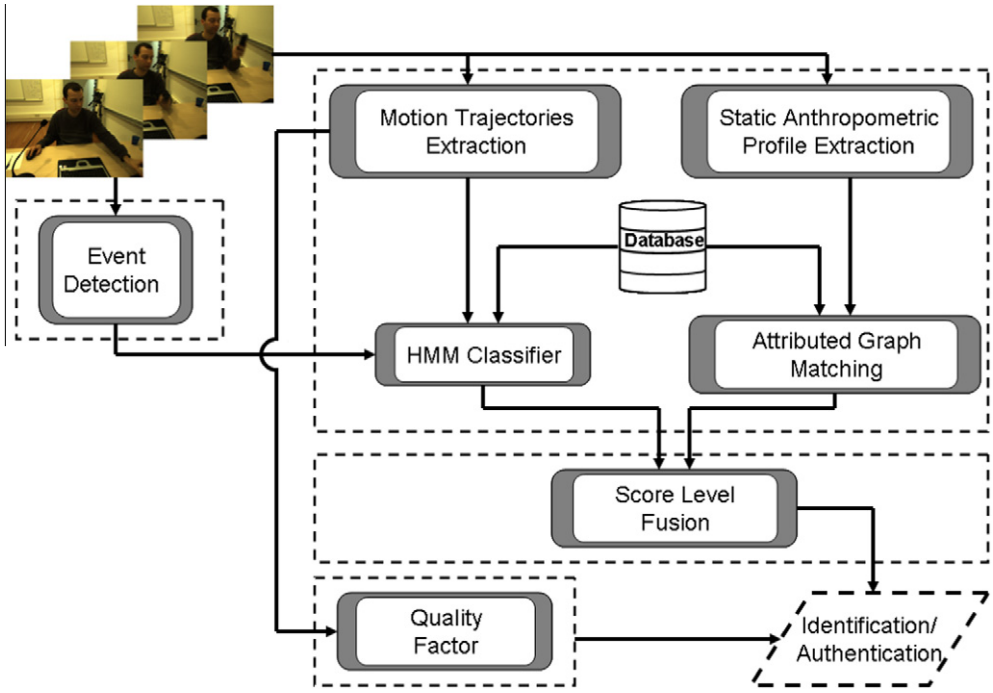

Fig. 1. User authentication system overview.

Fig. 2. MHI Samples: (a) Phone conversation – (b) typing pin on a wall-keyboard – (v) talking to microphone.

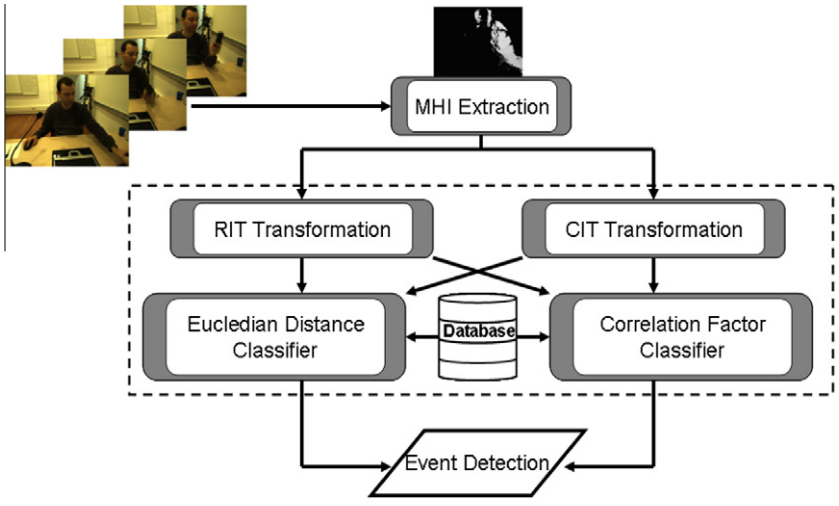

Fig. 3. Event detection system overview.

In particular, the RIT transform of a function $f(..,..)$ is defined as the integral of $f(..,..)$ along a line starting from the center of the image (x_0,y_0), which forms angle θ with the horizontal axis (Fig. 4/left). In our feature extraction method, the discrete form of the RIT transform is used, which computes the transform in steps of $\Delta\theta$ and is given by Eq. (2).

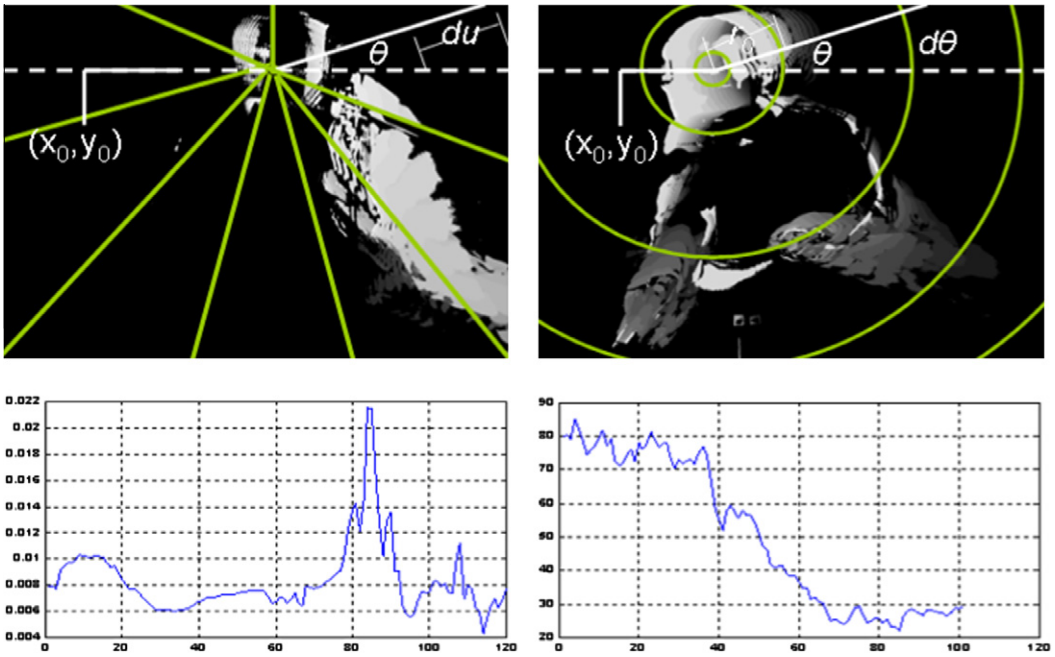

Fig. 4. Visual description of RIT (left) and CIT (right) for the activity "Phone Conversation" and "Talking to Mic. Panel" respectively.

$$RIT(t\Delta\theta) = \frac{1}{J}\sum_{j=1}^{J} MHI(x_0 + j\Delta u \cdot \cos(t\Delta\theta), y_0 + j\Delta u \cdot \sin(t\Delta\theta)) \quad (2)$$

for $t = 1, \ldots, T$ with $T = 360^o/\Delta\theta$, where $\Delta\theta$ and Δu are the constant step sizes of the distance u and angle θ and J is the number of the pixels that coincide with the line that has orientation R and are positioned between the center of the head and the end of the MHI in that direction.

In a similar manner, the CIT is defined as the integral of a function $f(..,..)$ along a circle curve with center (x_0,y_0) and radius ρ. Similar to the RIT transform, the discrete form of the CIT transform is used, as illustrated in Fig. 4/right, which is given by Eq. (3)

$$CIT(t\Delta\rho) = \frac{1}{T}\sum_{t=1}^{T} MHI(x_0 + k\Delta\rho \cdot \cos(t\Delta\theta), y_0 + k\Delta\rho \cdot \sin(t\Delta\theta)) \quad (3)$$

for $k = 1, \ldots, K$ with $T = 360^o/\Delta\theta$, where $\Delta\rho$ and $\Delta\theta$ are the constant step sizes of the radius and angle variables and finally $K\Delta\rho$ is the radius of the smallest circle that encloses the gray-scaled MHI (Fig. 4-right).

The database of supported activities consists of several sets of MHIs transformed according to RIT and CIT methods for each activity. Thus, an incoming transformed signal $\mathbf{x}$ is compared to a stored one $\mathbf{y}$ according to two separate classifiers; namely an Euclidian distance classifier, which deals with the transformed signals' absolute value (Eq. (4)) and a correlation factor classifier, which compares the fluctuations of the signals (Eq. (5)).

$$D_E = \sqrt{(\mathbf{x}-\mathbf{y})^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (4)$$

$$corr(x,y) = \rho_{\mathbf{x},\mathbf{y}} = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}} = \frac{\mathrm{E}((\mathbf{x}-\mu_{\mathbf{x}})(\mathbf{y}-\mu_{\mathbf{y}}))}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}} \quad (5)$$

The detected event is the one that has the most matches with the prototype MHIs from several subjects, stored in the database, according to a *majority voting* rule. Accordingly, an activity is considered to be performed within the successive appearance of a starting and an ending event. Moreover, an event is only then detected, when the returned scores from both classifiers exceed the experimentally selected thresholds, so as to minimize the false positives.

3.2. Dynamic motion trajectories

The core of the proposed authentication system is based on the dynamic motion tracking is presented in [29] and is briefly described in the following so as to make the paper self-contained. The user's movements are recorded by a stereo camera and the raw captured images are processed, in order to track the users head and hands via the successive application of filtering masks on the captured image. Specifically, a skin-color mask [42] combined with a motion-mask [40] can provide the location of the palms, while the head can be accurately tracked via a combination of a head detection algorithm [41] and a mean-shift object tracking algorithm [43]. The 2.5D information can be easily derived performing disparity estimation from the input stereoscopic image sequence.

The question that should be answered at this point is whether: "Does the position of the palm contain enough information to describe the movement of the whole arm or not". The answer is positive and it can be justified according to Lacquaniti and Soechting [44]. Specifically, they have proved that there is a clear dependency between the user's elbow and shoulder angular positions from trial to trial regardless the movement speed and the target orientation. The same idea about the consistency of reaching and grasping movements was presented later in [14] as well. Thus, it is rational to claim that the palm can be representative for the movement of the whole arm.

During the tracking session, a series of post-processing algorithms [29] applied on the raw tracked points, manage to extract smooth motion trajectories which are then used as biometric signatures. Finally, both the training and the identification procedure are implemented by a HMM. Specifically, a five state, left-to-right, fully connected HMM is trained from several enrollment sessions of the same user. This setting has been selected, given that the explored activities are not expected to return to a previous state. Similarly, it has been both intuitively found out and experimentally

proven that the five states optimally describe the corresponding motion trajectories. Accordingly, in the verification step the extracted features from a user are used as input to the stored HMM and the user is classified as client or impostor to the system.

3.3. Static anthropometric profile

A significant enhancement to the authentication performance of the system described in Section 3.2 can be achieved by exploiting the static anthropometric information of each user, i.e. a user-specific skeleton model. At this point, it should be clarified that the development of a new gesture recognition technique or the further improvement of an existing one is out of the scope of the current work. On the contrary, the goal is to exhibit the potential of static anthropometric features towards biometric recognition.

Thus, two state-of-the-art methods are utilized in the current section for the extraction of the users' static biometric profile. The first is described in [45], whereby hierarchical particle filtering is utilized towards the accurate shape adjustment of an articulated model to the user's body. The multi-camera environment requested by this approach is provided by two calibrated cameras: a stereo frontal camera and a usb-simple camera, which is placed on top of the user.

Alternatively to the aforementioned method, a faster and more accurate method has been lately released. The latter utilizes the PrimeSense® advanced depth-sensor in combination with the OpenNI [46] library. Thus, the human form is segmented automatically from the high precision depth image, while 48 essential points of the human body are simultaneously tracked in the 3*D* space.

The core of the OpenNI library is a machine learning algorithm that has been statistically trained by millions of images of people in different poses. The statistical compilation of all these data allows OpenNI to adjust the most appropriate skeleton model to each human body in terms of size and pose. The implemented methodology is covered by an international patent and is described in [47].

When comparing these two approaches, one could notice that in the current setting the particle filtering algorithm utilized in [45] requires ∼15 s for the processing of a single shot (1 shot $\equiv 1\,\frac{frame}{camera}$). However, it has been found out that an initial approximate manual annotation of the user's joints may significantly increase the performance of the algorithm with respect to the achieved accuracy.

On the other hand, the OpenNI algorithm exhibits much lower computational requirements (30 fps), with a slight decrease in accuracy. A comparison in terms of biometric recognition performance between the aforementioned methods, as well as their contribution to the carried out experiments follow in Section 4.2.

Once the location of all body's joints have been estimated, the extracted user's skeleton model can represented by an Attributed Relational Graph (ARG) $G = \{V, E, \{A\}, \{B\}\}$ [48], whereby V are the nodes, E the edges, and A and B the corresponding attributes, respectively. The nodes and the edges stand for the joints and the limbs of the actual body, respectively, as shown in Fig. 6. Attribute matrix **A** is not used, since no attributes for the joints are utilized in the current framework, while attribute matrix **B** corresponds to the lengths of the limbs (≡ distances between the adjacent joints).

3.4. Attributed graph matching

Possible noisy estimation of the limbs' lengths is compensated when calculating the mean value of each anthropometric attribute

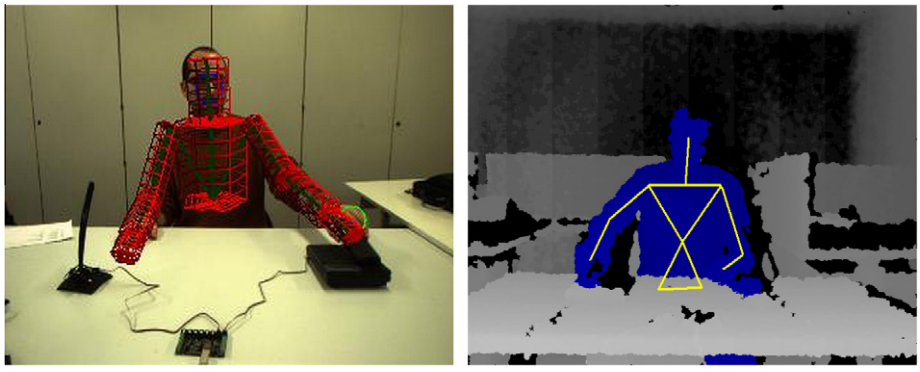

Fig. 5. Adjusted skeleton model based on: a) hierarchical filtering, b) OpenNI algorithms

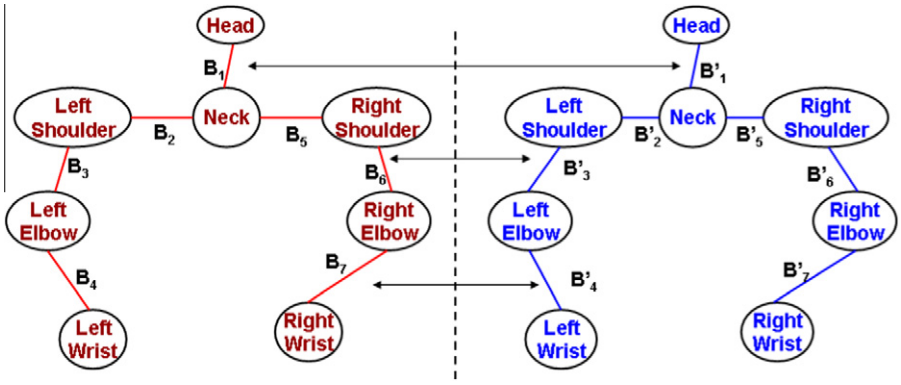

Fig. 6. Anthropometric Graphs' Comparison.

among several enrollment sessions. However, there are some cases, where partially connected anthropometric graphs may be generated. This may be due to either partial occlusions of specific limbs from other foreground objects or low confidence tracking (i.e. bad illumination). The Attributed Graph Matcher (AGM) based on Kronecker Graphs [48] has been utilized, whereby comparison between fully and partially connected graph is possible.

Let us assume two random anthropometric Graphs G and G' as shown below:

$$\begin{aligned} G &= \{V, E, \{\mathrm{B}\}_{i=1}^{n}\}, \text{ where } n := |V| \\ G' &= \left\{V', E', \{\mathrm{B}'\}_{i=1}^{n'}\right\}, \text{ where } n' := |V'| \end{aligned} \tag{6}$$

where B_k carries the lengths of the user's upper-body limbs.

The case of $n \neq n'$ indicates a Sub-Graph Matching (SGM), while $n = n'$ a Full-Graph Matching (FGM). In any case, Graph $\mathbf{G}$ is claimed to match to a sub-graph of $\mathbf{G}'$, if there exists an $n \times n'$ permutation sub-matrix $\mathbf{P}$ so that the following equation is fulfilled.

$$\mathbf{B_j} = \mathbf{P_0 B'_j P_0^T} + (\varepsilon \mathbf{M_j}), \text{ where } j = 1, \ldots, s \tag{7}$$

where $\mathbf{M_j}$ is an $n \times 1$ noise vector and ε is related to the noise power and is assumed to be independent of the indices i and j.

To accommodate inexactness in the modeling process due to noise, the AGM problem can be expressed as the combinatorial optimization problem of equation:

$$\epsilon = \min_{p} \left(\sum_{j=1}^{s} \mathbf{W_{j+r}} \|\mathbf{B_j} - \mathbf{PB'_j}\|^{\mathbf{q}} \right) \tag{8}$$

where $\|\cdot\|$ represents some norm $\mathbf{P} \in Per(n, n_0)$ denotes the set of all $n \times n_0$ permutation submatrices and $\{\mathbf{W_i}\}_{k=1}^{r+s}\}$ is a set of weights satisfying $0 \leqslant \mathbf{W_k} \leqslant 1$, $k = 1, \ldots, r+s$ and $\sum_{k=1}^{r+s} \mathbf{W_k} = 1$.

In this respect, the minimum error ϵ stands for a metric for the similarity between the graphs under comparison.

3.5. Ergonomy – quality factor

In order to fulfill the repeatability/reproducibility requirement (see Section 1.3) the same or almost the same environmental conditions should remain stable among different sessions. Moreover, the stylish and behavioral analysis of a person's movements always refers to a relaxed state. Otherwise, unwanted artifacts may appear, which will act as noise to the measurements. In the following, a method based on ergonomical studies is presented, which can handle the "extreme" cases of movements.

3.5.1. Ergonomic spheres

Due to restrictions set by the structure of the human body, it is easy to understand that there are regions around the human, where the movement of the hands is more convenient than in other regions. These assumptions have been scientifically formulated in [49]. Specifically, it has been proven that the area in front of a seated human can be divided in three different spheres, according to the easiness with which the user can reach an object within certain regions (Fig. 7). It is suggested that the darkly gray area is the one where the user moves most convenient and is thus called the "convenient zone". On the contrary, the light gray area indicates the "kinetosphere", whereby the user has to stretch or to bend his body in order to reach something. The white areas on Fig. 7 are out of reach for the user.

Thus, it can be assumed that the user performs more relaxed movements within the "convenient zone" than in the "kinetosphere". During run-time, it can be claimed that the movements within the "convenient zone" reveal more information about the user's behavioral response, since they are performed under no pressure or with force. On the other hand, the movements within the "kinetosphere" can be considered as forced movements. Thus, the ergonomic zones taken into account are dependent on the distance between the user's torso and the interaction objects.

In this respect, an important metric about the quality and the evaluation of the extracted signature is proposed and is defined in Eq. (9) as the product of the tracking quality factor f_q (Eq. (10)), enhanced by a user-object distance factor $b(0 \leqslant b \leqslant 1)$, which changes over the human ergonomic spheres (Eq. (11)).

$$f_{q,final} = b \cdot f_q \tag{9}$$

$$f_q = 1 - \frac{N_{missHead} + N_{missRHand} + N_{missLHand}}{3N_{frames}} \tag{10}$$

where $N_{missHead}$, $N_{missRHand}$, $N_{missLHand}$ are the amount of frames in which the Head, the right and the left Hand were not detected, respectively. N_{frames} is the total number of frames of the sequence.

$$b = \begin{cases} 0.1 \cdot d_{torso,object} + 0.5, & \text{if } d_{torso,object} < 5cm \\ 1, & \text{if } 5cm \leqslant d_{torso,object} \leqslant 35cm \\ -0.02 \cdot d_{torso,object} + 1.7, & \text{if } d_{torso,object} > 35cm \end{cases} \tag{11}$$

The lowest the quality factor the less probable the extracted dynamic features to contain valuable biometric information for authentication. Accordingly, if $f_{q,final} \leqslant 0.5$ the extracted features are discarded and no authentication process takes place.

The quality factor can be used in favor of the authentication rate of the system as follows: Forced movements that include the stretching of the user are inherently different both in style and in

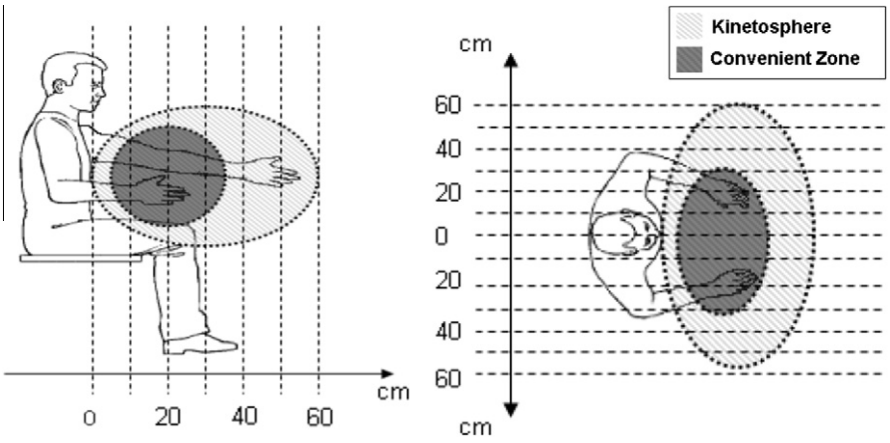

Fig. 7. Human convenience zones on a table.

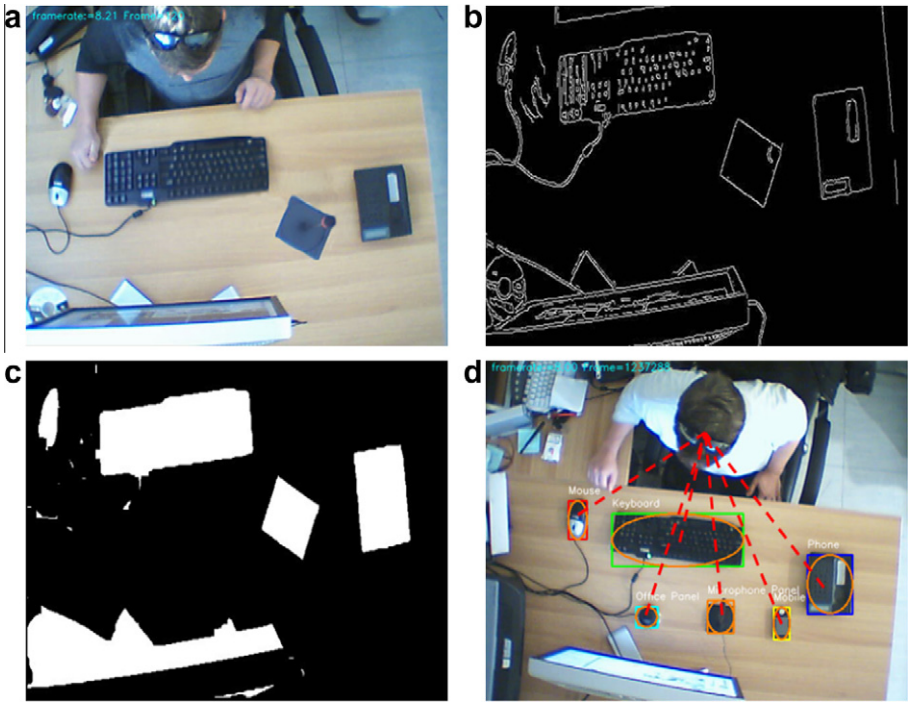

Fig. 8. Object detection: (a) Top camera view, (b) contour extraction (c) objects' area detection, and (d) tagging of objects.

type. Thus, no authentication potential is expected to be found in such movements, given that in the current study the user is expected to act under regular, relaxed conditions, similar to the ones during the enrollment session. In this respect, the quality factor described above contributes to the implicit detection of such movements, in order to be excluded from classification.

3.5.2. Torso – object distance estimation

The proposed multicamera environmental setting (Fig. 5) consists of two calibrated cameras (i.e. a frontal stereo-camera and a top monocular camera). In order to calculate the distance $d_{torso,object}$, both the torso and each object have to be first detected on the recording setting. Given that the head position is detected as described in Section 3.2, the underlying body part refers to the user's torso. On the other hand, each object can be detected by the top camera as shown in Fig. 8 and . Generally, objects are coarsely described in a rotation-invariant way based on their contours (Fig. 8b). Specifically, each object is described by its aspect ratio, the area it occupies and its color.

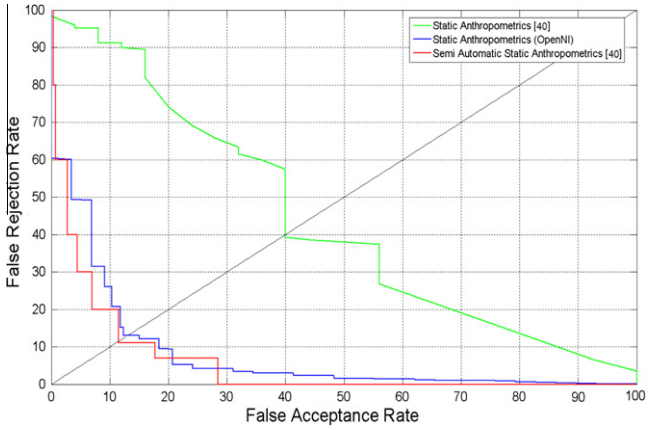

Fig. 9. ROC Curves for the static biometric traits.

Since the two cameras are calibrated with each other, the distance $d_{torso,object}$ can be easily calculated as illustrated by the red dotted lines shown in Fig. 8d.

3.6. Fusion

In order to combine the results from the two different biometric traits, namely the dynamic and the static one, a score-level fusion algorithm has been utilized. Specifically, the fusion is performed by a support vector machine (SVM) classifier that bases on a gaussian kernel with a width value of 0.01. The trade-off factor between training error and margin was set at 100,000, while all input score-data have undergone a "$min - max$" normalization. The training of the SVM has been performed on a $19 - subjects$ custom-dataset, which is described in paragraph Section 4.1.3.

4. Databases and results

The proposed framework has been evaluated in the context of the following verification scenarios, namely "a short phone conversation" (scenario A) and "talking to a microphone panel" (scenario B). In both cases, the verification results based on the static anthropometric profile of each user, the dynamic motion trajectories and their fusion are presented.

4.1. Databases

For the evaluation of the current framework the following three databases have been utilized:

4.1.1. ACTIBIO database

This database was captured in an ambient intelligence indoor environment and is extensively described in [29]. More precisely, the current, manually annotated database consists of 29 subjects, performing a series of everyday office activities, i.e. a phone conversation, typing, talking to a microphone panel, drinking water, etc., with no special protocol. Each subject has performed

8 repetitions in total, equally split in two sessions. Among the five cameras which have been recording each user from different angles, only the recordings from a frontal stereo camera and a top monocular camera have been used for the current work.

4.1.2. Anthropometric database

In order to acquire the appropriate recordings for the OpenNI algorithms, a custom dataset has been recorded by the PrimeSense® camera sensor. This dataset consists of 29 subjects performing the activities indicated in the scenarios A and B in 3 repetitions. 200 frames from each of the first two repetitions have been used for extracting the user's anthropometric profile, by which each user was registered to the database. Similarly, the anthropometric profile that has been used for authentication was formed by averaging the results of 200 sequential frames from the third repetition.

4.1.3. Fusion training database

The training of the SVM has been performed on the data acquired from a third, 19-*subjects* custom-dataset. Specifically, this dataset has been recorded in a multi-camera scenario, which consisted of a frontal stereo camera and a top usb-camera. All subjects performed the activities indicated in the two aforementioned scenarios in 2 sessions of 5 repetitions in total.

4.2. Results

In the following, the performance of the main modules of the proposed framework (i.e. the activity detection module and the activity-related authentication module) is evaluated, during the performance of several activities.

4.2.1. Activity detection performance

The performance of the proposed activity detection framework (Section 3.2) exhibited high accuracy, as described in the confusion matrix of Table 1. The performed experiment forced a simultaneous search for the detection of the four supported activities. Table 1 presents the high detection potential of the proposed approach in realistic unobtrusive conditions.

Activities with high motion content in close areas of the frame, i.e. glass and phone are both brought towards the user's head, are most likely to be mismatched. On the other hand, activities exhibiting high motion variance, i.e. the users picks the phone with the left hand and speaks to the microphone on his right side, are highly probable to be correctly detected.

4.2.2. Authentication performance

Considerable improvements in the authentication performance compared to [29] have been observed when augmenting the motion trajectory based algorithm with the static anthropometric information. In the new experiments that have been carried out, the authentication performance of the dynamic traits from the two scenarios are illustrated with the red line in Figs. 10 and 11. The reader can easily notice that *scenario B* exhibits a higher authentication rate (EER = 10.32%) compared to the *scenario A* (EER = 16.7%). This can be explained by the fact that during a short phone conversation the user's head remains almost fixed at the same position. Thus, the recognition capacity of the movement is mainly concentrated in the movement of the hand. On the other hand, the second scenario required the leaning of the user towards the microphone. In this respect, both the head and the user's hand covered a significant distance. Thus, more valuable biometric information is potentially encoded in the second movement.

Table 1
Activity detection confusion matrix.

Events	Phone (%)	Panel (%)	Microphone (%)	Drinking (%)
Phone	93.1	0	0	6.9
Panel	0	89.7	10.3	0
Microphone	0	10.3	86.2	3.44
Drinking	13.8	0	0	86.2

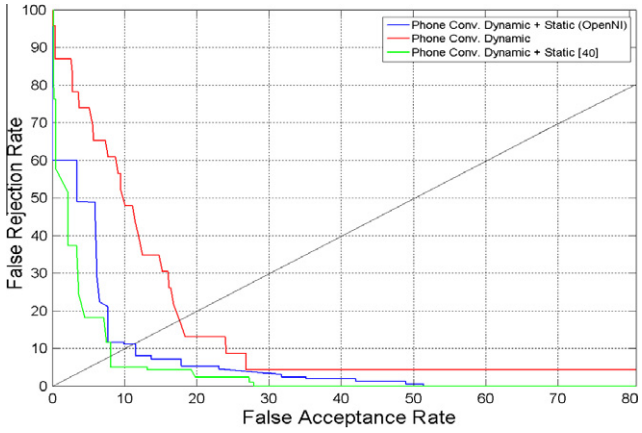

Fig. 10. Scenario A – ROC Curves for the fused scores.

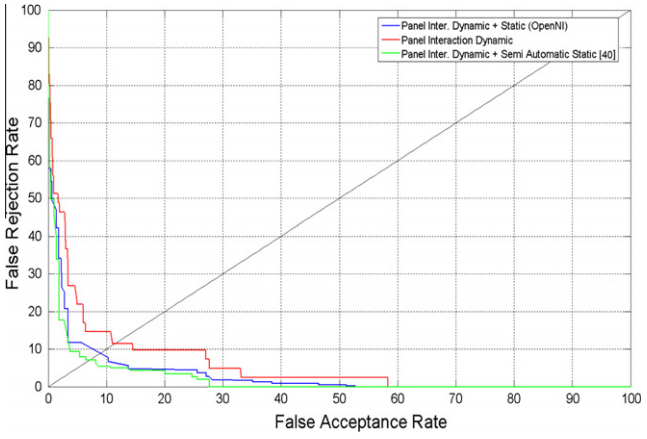

Fig. 11. Scenario B – ROC Curves for the fused scores.

Fig. 9 exhibits the authentication potential of the static anthropometric characteristics, as they are processed based on the approaches described in Section 3.3. The blue line depicts the authentication potential of the particle filtering method[45]. Although, the tracking of the user seems to be accurate, the reader can notice that the method fails to provide significant authentication potential with respect to the individual's anthropometric information. On the contrary, when the initialization of the algorithm is augmented by manual annotation of the user's joints, the algorithm exhibits high robustness. In this semi-automatic version of the algorithm the authentication rate lies at 11.3%.

The superiority of the proposed method, which utilizes the Primesense sensor with the OpenNI algorithms is clear in both computational time and accuracy. Specifically, the fully automatic method implemented be the OpenNI algorithms achieves an authentication rate score of 13.23%, which is very close to the performance of the semi-automatic particle filtering method described above. Moreover, the computation time needed has significantly decreased, and thus a real-time anthropometric profile extraction is possible.

As expected, when combining both static and dynamic extracted information the authentication performance of the system

Table 2
Scenario A – authentication performance (EER).

	Dynamic	Static (%)		Fusion (%)		Fus. and Ergon. (%)	
		[45]	[46]	[45]	[46]	[45]	[46]
Scenario A	16.7	11.3	13.23	8.3	10.8	7.9	10.1
Scenario B	10.32	11.3	13.23	7.2	9.12	6.7	8.4

improves further. Specifically, the fusion performed by the SVM presented in Section 3.6 achieved an EER score of 8.3% in scenario A and 7.2% in scenario B, when the fully automatic has been utilized. The EER scores are even lower in the case the semi-automatic particle filtering method has been utilized as shown in Table 2.

The EERs are summarized in Table 2, whereby the improvements of the proposed ergonomy-based quality factor (Sections 3.5.1 and 3.5.2) are included. Specifically, it is shown that the EER when the ergonomy restrictions are applied falls with a mean value of 0.6% in the both experimental scenarios. This improvement stems from the fact that specific repetitions have been excluded from evaluation in the authentication step, since they exhibited low ergonomic confidence. Thus, a reduced false rejection rate has been achieved.

5. Conclusions

In this paper an extension to an activity-related, unobtrusive authentication framework has been presented, that is related to activity-related biometrics and includes both the dynamic and the static characteristics derived when performing everyday activities. The proposed framework can be expanded to various activities, which include the reaching, grasping or interacting with an object in the vicinity of a user.

The system is triggered by a robust event detection algorithm, while the quality of the extracted features is verified with respect to both, the accuracy of the tracking algorithm and the ergonomy of the setting. The proposed framework is seen experimentally to provide very promising verification rates in real time. Moreover, the proposed static anthropometric profile is seen to have a significant contribution to the overall authentication capacity. Last, taking also into account that no hard constraints have been forced during the capture of the input signals, the proposed approach makes a step forward in the context of the very challenging problem of unobtrusive on-the-move-biometry.

Acknowledgment

This work was supported by the EU funded ACTIBIO ICT STREP (FP7-215372).

References

[1] F.A. Qazi, A survey of biometric authentication systems, Secur. Manage. (2004) 61–67.
[2] Q. Xiao, Security issues in biometric authentication, Information Assurance Workshop, IAW 2005 (2005) 8–13.
[3] D. Kawanaka, T. Okatani, K. Deguchi, HHMM based recognition of human activity, IEICE Trans. Inf. Syst. (Inst. Electron. Inf. Commun. Eng.) 7 (2006) 2180–2185.
[4] N. Zouba, F. Bremond, M. Thonnat, V.T. Vu, Multi-sensors analysis for everyday activity monitoring, in: 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications (SETIT2007), 2007.
[5] Q. Dong, Y. Wu, Z. Hu, Pointwise Motion Image (PMI): a novel motion representation and its applications to abnormality detection and behavior recognition, IEEE Trans. Circ. Syst. Video Technol. 19 (3) (2009) 407–416, doi:10.1109/TCSVT.2009.2013503.
[6] W. Lin, M.-T. Sun, R. Poovendran, Z. Zhang, Activity recognition using a combination of category components and local models for video surveillance, IEEE Trans. Circ. Syst. Video Technol. 18 (8) (2008) 1128–1139, doi:10.1109/TCSVT.2008.927111.
[7] L. Benedikt, D. Cosker, P. Rosin, D. Marshall, Assessing the uniqueness and permanence of facial actions for use in biometric applications, IEEE Trans. Syst., Man, Cybern. A, Syst., Humans 40 (3) (2010) 449–460, doi:10.1109/TSMCA.2010.2041656.
[8] A.P. Glascock, D.M. Kutzik, Behavioral telemedicine: a new approach to the continuous nonintrusive monitoring of activities of daily living, Telemed. J. 6 (1) (2004) 33–44, doi:10.1089/107830200311833.
[9] J. Lester, T. Choudhury, N. Kern, G. Borriello, B. Hannaford, A hybrid discriminative/generative approach for modeling human activities, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2005, pp. 766–772. doi:10.1.1.77.5776.
[10] E. Tapia, S.S. Intille, K. Larson, Activity recognition in the home using simple and ubiquitous sensors, in: A.F. Mattern (Ed.), Proceedings of Pervasive, Springer, Berlin Heidelberg, 2004, pp. 158–175.
[11] F. Fusier, V. Valentin, F. Brémond, M. Thonnat, M. Borg, D. Thirde, J. Ferryman, Video understanding for complex activity recognition, Mach. Vis. Appl. 18 (3) (2007) 167–188, doi:10.1007/s00138-006-0054-y.
[12] L. Wang, D. Suter, Learning and matching of dynamic shape manifolds for human action recognition, IEEE Trans. Image Process. 16 (6) (2007) 1646–1661, doi:10.1109/TIP.2007.896661.
[13] I. Junejo, E. Dexter, I. Laptev, P. Perez, View-independent action recognition from temporal self-similarities, IEEE Trans. Pattern Anal. Mach. Intell. 33 (1) (2011) 172–185, doi:10.1109/TPAMI.2010.68.
[14] D.A. Rosenbaum, R.J. Meulenbroek, J. Vaughan, C. Jansen, Posture-based motion planning: applications to grasping, Phsychol. Rev. 108 (4) (2001) 709–734.
[15] A.K. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition, IEEE Trans. Circ. Syst. Video Technol. 14 (1) (2004) 4–20, doi:10.1109/TCSVT.2003.818349.
[16] M. Garris, E. Tabassi, C. Wilson, NIST fingerprint evaluations and developments, Proc. IEEE 94 (11) (2006) 1915–1926, doi:10.1109/JPROC.2006.885130.
[17] N. Fox, R. Gross, J. Cohn, R. Reilly, Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts, IEEE Trans. Multimedia 9 (4) (2007) 701–714, doi:10.1109/TMM.2007.893339.
[18] N. Schmid, M. Ketkar, H. Singh, B. Cukic, Performance analysis of iris-based identification system at the matching score level, IEEE Trans. Inform. Forensics Secur. 2 (1) (2006) 154–168, doi:10.1109/TIFS.2006.873603.
[19] G. Zheng, C.-J. Wang, T. Boult, Application of projective invariants in hand geometry biometrics, IEEE Trans. Inform. Forensics Secur. 2 (4) (2007) 758–768, doi:10.1109/TIFS.2007.908239.
[20] A. Jain, J. Feng, Latent palmprint matching, IEEE Pattern Anal. Mach. Intell. 31 (2009) 1032–1047, doi:10.1109/TPAMI.2008.242.
[21] A. Kale, N. Cuntoor, R. Chellappa, A framework for activity-specific human identification, in: IEEE Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 4, 2002, pp. 3660–3663.
[22] K. Delac, M. Grgic, A survey of biometric recognition methods, in: Proceedings Elmar 2004. in: 46th International Symposium Electronics in Marine, 2004, pp. 184–193.
[23] H. Junker, J. Ward, P. Lukowicz, G. Tröster, User Activity Related Data Sets for Context Recognition, in: Proceedings of Workshop on 'Benchmarks and a Database for Context Recognition', 2004.
[24] A. Kale, A. Sundaresan, A. Rajagopalan, N.P. Cuntoor, A.K. Roy-Chowdhury, V. Kruger, R. Chellappa, Identification of humans using gait, IEEE Trans. Image Process. 13 (2004) 1163–1173.
[25] A. Hadid, M. Pietikäinen, S.Z. Li2, Learning Personal Specific Facial Dynamics for Face Recognition from Videos, Springer, Berlin/Heidelberg, 2007. pp. 1–15. doi:10.1007/978-3-540-75690-3_1.
[26] D. Ioannidis, D. Tzovaras, K. Moustakas, Gait identification using the 3D protrusion transform, in: IEEE International Conference on Image Processing (ICIP), San Antonio, 2007, pp. I-349 – I-352. doi:10.1109/ICIP.2007.4378963.
[27] M. Goffredo, I. Bouchrika, J.N. Carter, M.S. Nixon, Performance analysis for automated gait extraction and recognition in multi-camera surveillance, Multimedia Tools and Applications, doi:10.1007/s11042-009-0378-5.
[28] M.S. Nixon, J.N. Carter, J.M. Nash, P.S. Huang, D. Cunado, S.V. Stevenage, Automatic gait recognition, IEE Colloq. Mot. Anal. Track. (1999) 3/1–3/6.
[29] A. Drosou, K. Moustakas, D. Ioannidis, D. Tzovaras, On the potential of activity-related recognition, in: The International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2010), 2010.
[30] T. Moeslund, A. Hilton, V. Krueger, A survey of advances in vision-based human motion capture and analysis, Comput. Vis. Image Underst. 104 (2) (2006) 90–126.
[31] R. Plankers, P. Fua, Articulated soft objects for multiview shape and motion capture, IEEE Trans. Pattern Anal. Mach. Intell. 25 (9) (2003) 1182–1187, doi:10.1109/TPAMI.2003.1227995.
[32] J. Ziegler, K. Nickel, R. Stiefelhagen, Tracking of the articulated upper body on multi-view stereo image sequences, in: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06), 2006, pp. 774–781. doi:10.1109/CVPR.2006.313.
[33] A. Micilotta, E.-j. S.Ong, R. Bowden, Real-time upper body detection and 3d pose estimation in monoscopic images, in: European Conference on Computer Vision, 2006, pp. 139–150.
[34] M. Holte, T. Moeslund, P. Fihl, View-invariant gesture recognition using 3D optical flow and harmonic motion context, Comput. Vis. Image Underst. 114 (12) (2010) 1353–1361. doi:http://dx.doi.org/10.1016/j.cviu.2010.07.012.
[35] M. Turvey, Perceiving, Acting, and Knowing: Toward an Ecological Psychology, Lawrence Erlbaum, New Jersey, 1977.
[36] S.R. Goodman, G. Gottlieb, Analysis of kinematic invariances of multijoint reaching movement, Biol. Cybern. 73 (1995) 311–322.

[37] T. Flash, N. Hogan, The coordination of arm movements: an experimentally confirmed mathematical model, J. Neurosci. 5 (7) (1985) 1688–1703.
[38] Y.K. Uno, R.M. Suzuki, Formation and control of optimal trajectory in human multijoint arm movement – minimum torque-change model, Biol. Cybern. 61 (2) (1989) 89–101.
[39] A. Drosou, K. Moustakas, D. Tzovaras, Event-based unobtrusive authentication using multi-view image sequences, in: Proceedings of ACM Multimedia/Artemis Workshop ARTEMIS 10, Florence, 2010, pp. 69–74.
[40] A. Bobick, J. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (3) (2001) 257–267, doi:10.1109/34.910878.
[41] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple, in: IEEE Proceedings of Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2001, pp. 511–518. doi:10.1109/CVPR.2001.990517.
[42] G. Gomez, E.F. Morales, Automatic feature construction and a simple rule induction algorithm for skin detection, in: Proceedings of the ICML Workshop on Machine Learning in Computer Vision (MLCV), 2002, pp. 31–38.
[43] D.C.V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, in: IEEE Proceeings of Computer Vision and Pattern Recognition 2007 (CVPR), vol. 2, 2000, pp. 142–149. doi:10.1109/CVPR.2000.854761.
[44] F. Lacquaniti, J.F. Soechting, Coordination of arm and wrist motion during a reaching task, J. Neurosci.: Off. J. Soc. Neurosci. 2 (4) (1982) 399–408.
[45] M. Alcoverro, J.R. Casas, M. Pardas, Skeleton and Shape Adjustment and Tracking in Multicamera Environments, vol. 6169, Mallorca, 2010, pp. 88–97. doi:10.1007/978-3-642-14061-7_9.
[46] L. PrimeSense, W. Garage, S. Production, Natural Interaction, 2010. <http://www.openni.org/>.
[47] M.G. Ryan, Visual Target Tracking. US patent 20100195869, issued Jan.30.2009.
[48] B.V. Wyk, M.V. Wyk, Kronecker product graph matching, Pattern Recogn. 36 (9) (2003) 2019–2030, doi:10.1016/S0031-3203(03)00009-8.
[49] A. Toney, B.H. Thomas, Considering reach in tangible and table top design. In: 1st IEEE International Workshop on Horizontal Interactive Human–Computer Systems, 2006, pp. 2. doi:10.1109/TABLETOP.2006.9.