A Human Body Analysis System

Vincent Girondel, Laurent Bonnaud and Alice Caplier
Laboratoire des Images et des Signaux (LIS), INPG, France
http://www.lis.inpg.fr

***Abstract*—** This paper describes a system for human body analysis (segmentation, tracking, face/hands localisation, posture recognition) from a single view that is fast and completely automatic. The system first extracts low-level data and uses part of the data for high-level interpretation. It can detect and track several people even if they merge or are completely occluded by another person from the camera's point of view. For the high-level interpretation step, static posture recognition is performed using a belief theory-based classifier. The belief theory is considered here as a new approach for performing posture recognition and classification using imprecise and/or conflicting data. Four different static postures are considered: standing, sitting, squatting and lying.

The aim of this paper is to give a global view and an evaluation of the performances of the entire system and to describe in detail each of its processing steps, whereas our previous publications focused on a single part of the system. The efficiency and the limits of the system have been highlighted on a database of more than fifty video sequences where a dozen different individuals appear. This system allows real-time processing and aims at monitoring elderly people in video surveillance applications or at the mixing of real and virtual worlds in ambient intelligence systems.

***Keywords*—**belief theory, face detection, human body analysis, human posture recognition, real-time processing, skin detection.

I. INTRODUCTION

HUMAN motion analysis is an important area of research in computer vision devoted to detecting, tracking and understanding people's physical behaviour. This strong interest is driven by a wide spectrum of applications in various areas such as smart video surveillance [1], interactive virtual reality systems [2, 3], advanced and perceptual human-computer interfaces (HCI) [4], model-based coding [5], content-based video storage and retrieval [6], sports performances analysis and enhancement [7], clinical studies [8], smart rooms and ambient intelligence systems [9, 10] etc. The "looking at people" research field has recently received a lot of attention [11, 12, 13, 14, 15, 16]. Here, the considered applications are video surveillance and smart rooms with advanced HCIs.

Video surveillance covers applications where people are being tracked and monitored for particular actions. The demand for smart video surveillance systems comes from the existence of security-sensitive areas such as banks, department stores, parking lots etc. Surveillance cameras video streams are often stored in video archives or recorded on tapes. Most of the time, these video streams are only used "after the fact" mainly as an identification tool. The fact that the camera is an active sensor and a real-time processing media is therefore sometimes unused. The need is the real-time video analysis of sensitive places in order to alert the police of a burglary in progress, or of the suspicious presence of a person wandering for a long time in a parking lot. As well as obvious security applications, smart video surveillance is also used to measure and control the traffic flow, compile consumer demographics in shopping malls, monitor elderly people in hospitals or at home etc.

W^4: "Who? When? Where? What?" is a real-time visual surveillance system for detecting and tracking people and monitoring their activities in an outdoor environment [1]. It operates on monocular grey scale or on infrared video sequences. It makes no use of colour cues, instead it uses appearance models employing a combination of shape analysis and tracking to locate people and their body parts (head, hands, feet, torso) and track them even under occlusions. Although the system succeeds in tracking multiple people in an outdoor complex environment, the cardboard model used to predict body posture and activity is restricted to upright people, i.e. recognised actions are, for example, standing, walking or running. The DARPA VSAM project lead to a system for video-based surveillance [17]. Using multiple cameras, it classifies and tracks multiple people and vehicles. Using a star skeletonization procedure for people, it succeeds in determining the gait and posture of a moving human being, classifying its motion between walking and running. As this system is designed to track vehicles or people, human subjects are not big enough in the frame, so the individual body components can not be reliably detected. Therefore the recognition of human activities is restricted to gait analysis. In [18], an automated visual surveillance system that can classify human activities and detect suspicious events in a scene is described. This real-time system detects people in a corridor, tracks them and uses dynamic information to recognise their activities. Using a set of discrete and previously trained Hidden Markov Models (HMMs), it manages to classify people entering or exiting a room, and even mock break-in attempts. As there are many other possible activities in a corridor, for instance speaking with another person, picking up an object on the ground, or even lacing shoes squatting near a door, the system has a high false alarm rate.

For advanced HCIs, the next generation will be multi modal, integrating the analysis and recognition of human body postures and actions as well as gaze direction, speech and facial expressions analysis. The final aim of [4] is to develop human-computer interfaces that react in a similar way to a communication between human beings. Smart rooms and ambient intelligence systems offer the possibility of mixing real and virtual worlds in mixed reality applications [3]. People entering a camera's field of view are placed into a virtual environment. Then they can interact with the environment, with its virtual objects and with other people (using another instance of the system), by their behaviour (gestures, postures or actions) or by another media (for instance speech).

Pfinder is a real-time system designed to track a single human in an indoor environment and understand its physical behaviour [2]. It models the human body and its parts using small blobs with numerous characteristics (position, colour, shape etc.). The background and the human body are modelled with Gaussian distributions and the human body pixels are classified as belonging to particular body parts using the log-likelihood measure. Nevertheless, the presence of other people in the scene will affect the system as it is designed for a single person. Pfinder has been used to explore several different HCIs applications. For instance, in ALIVE and SURVIVE (respectively [9] and [10]), a 3D virtual game environment can be controlled and navigated through by the user gestures and position.

In this paper, we present a system that can automatically detect and track several people, their faces and hands and recognise in real-time four static human body postures (standing, sitting, squatting and lying). Whereas our previous publications focused on a single part of the system, here the entire system is described in detail and both an evaluation of the performances and a discussion are given. Low-level data are extracted using dynamic video sequence analysis. Then, depending on the desired application, part or all of these data can be used for human behaviour high-level recognition and interpretation. For instance, static posture recognition is performed by data fusion using the belief theory. The belief theory is considered here as a new approach for performing posture recognition.

Overview

Overview of the paper

Sections II to V present the low-level data extraction processing steps: 2D segmentation of people (II), basic temporal tracking (III), face and hands localisation (IV) and Kalman filtering-based tracking (V). Section VI illustrates an example of high-level human behaviour interpretation, dealing with static posture recognition. Finally section VII concludes the paper, discusses the results of the system and gives some perspectives.

Overview of the system

As processing has to be close to real-time, the system has some constraints in order to design low-complexity algorithms. Moreover, with respect to the considered applications, they are not so restrictive. The general constraints, necessary for all processing steps, are:

1. The environment is filmed by **one static camera**.
2. People are the only both **big** and **mobile objects**.
3. Each person enters the scene **alone**.

The constraint n°1 comes from the segmentation processing step, as it is based on a background removal algorithm. The constraints n°2 and n°3 follow from the aim of the system to analyse and interpret human behaviour. They are assumed to facilitate the tracking, the face and hands localisation and the static posture recognition processing steps.

Fig. 1 gives an overview of the system. On the left side are presented the processing steps and on the right side the resulting data. Fig. 2 illustrates the processing steps.

Glossary

- FRBB: Face Rectangular Bounding Box
- FPRBB: Face Predicted Rectangular Bounding Box
- FERBB: Face Estimated Rectangular Bounding Box
- ID: IDentification number
- PPRBB: Person Predicted Rectangular Bounding Box
- PERBB: Person Estimated Rectangular Bounding Box
- SPAB: Segmentation Principal Axes Box
- SRBB: Segmentation Rectangular Bounding Box

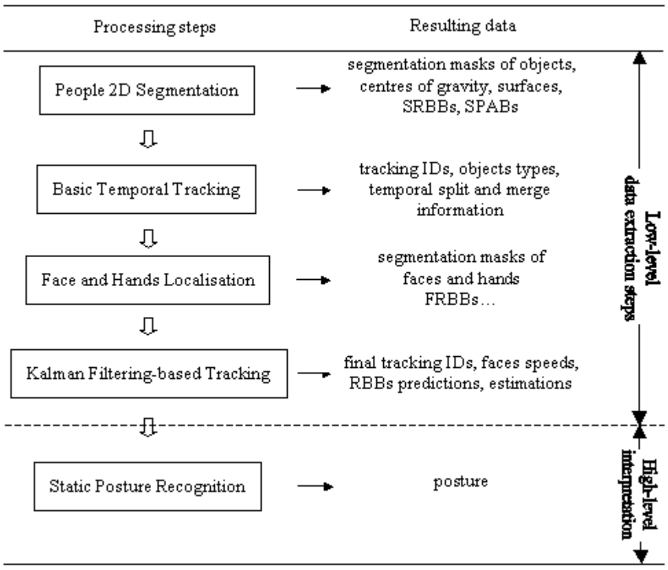


Fig. 1. Overview of the system.

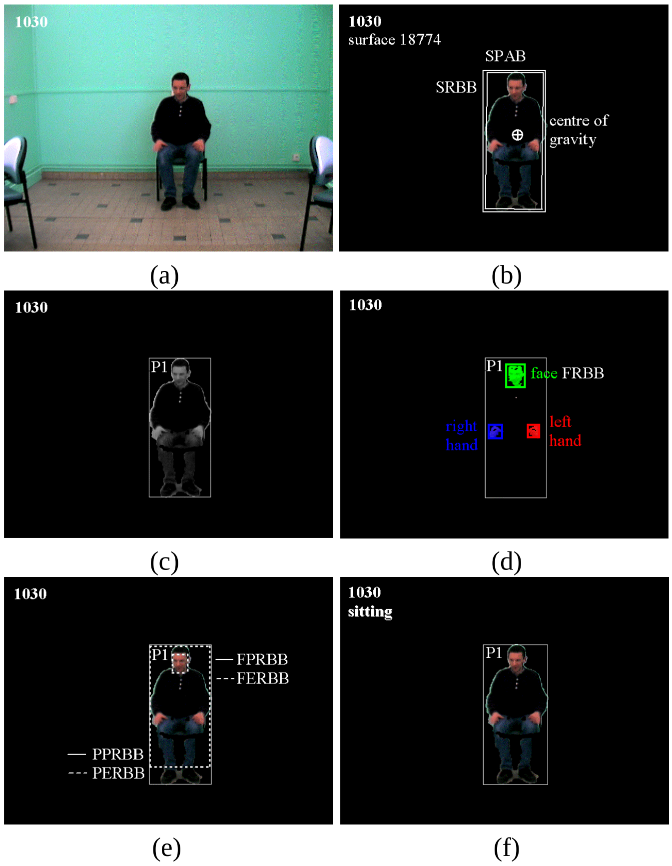


Fig. 2. Example of system processing steps. (a) original frame, (b) people 2D segmentation, (c) basic temporal tracking, (d) face and hands localisation, (e) Kalman filtering-based tracking and (f) static posture recognition.

II. PEOPLE 2D SEGMENTATION

Like most vision-based systems whose aim is the analysis of human motion, the first step is the extraction of people present in the scene. Considering people moving in an unknown environment, this extraction is a difficult task [19]. It is also a significant issue since all the subsequent steps such as tracking, skin detection and posture or action recognition are greatly dependent on it.

A. Our approach

When using a static camera, two main approaches have been considered. On the one hand, only consecutive frames differences are used [20, 21, 22], but one of the major drawbacks is that no temporal changes occur on the overlapped region of moving objects especially if they are low textured. Moreover, if the objects stop, they are no more detected. As a result, segmented video objects may be incomplete. On the other hand, only a difference with a reference frame is used [23, 24, 25]. It gives the whole video object area even if the object is low textured or stops. But the main problem is the building and updating of the reference frame. In this paper, moving people segmentation is done using the Markov Random Field (MRF) based motion detection algorithm developed in [26] and improved in [27]. The MRF modelling involves consecutive frame differences and a reference frame in a unified way. Moreover the reference frame can be built even if the scene is not empty.

Fig. 3 summarises the 2D segmentation processing step.

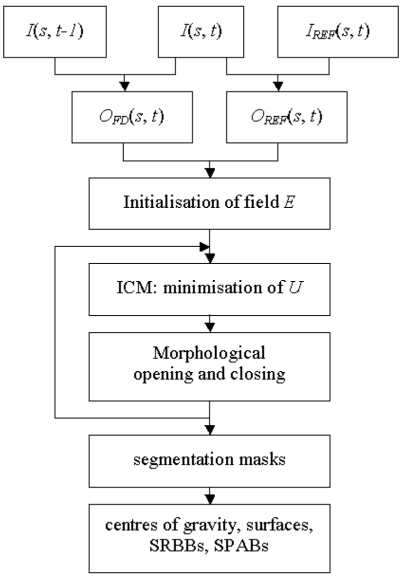


Fig. 3. Scheme of the people 2D segmentation processing step.

B. Labels and observations

Motion detection is a binary labelling problem which aims at attributing to each pixel or "site" $s = (x, y)$ of frame I at time t one of the two possible labels:

$$e(x, y, t) = e(s, t) = \begin{cases} obj \text{ if } s \text{ belongs to a person} \\ bg \text{ if } s \text{ belongs to the background} \end{cases}$$

$e = \{e(s, t), s \in I\}$ represents one particular realization (at time t) of the label field E. Additionally, we define $\{e\}$ as the set of possible realizations of field E.

With the constraint n°1 of the system, motion information is closely related to temporal changes of the intensity function $I(s, t)$ and to the changes between the current frame $I(s, t)$ and a reference frame $I_{REF}(s, t)$ which represents the static background without any moving people. Therefore, two observations are defined:

- an observation O_{FD} coming from consecutive frame differences:

$$o_{FD}(s, t) = |I(s, t) - I(s, t-1)|$$

- an observation O_{REF} coming from a reference frame:

$$o_{REF}(s, t) = |I(s, t) - I_{REF}(s, t)|$$

$o_{FD} = \{o_{FD}(s, t), s \in I\}$ and $o_{REF} = \{o_{REF}(s, t), s \in I\}$ represent one particular realization (at time t) of the observation fields O_{FD} and O_{REF} respectively.

To find the most probable configuration of field E given fields O_{FD} and O_{REF}, we use the MAP criterion and look for $e \in \{e\}$ such that ($Pr[\,]$ denotes probability):

$$Pr[E = e/O_{FD} = o_{FD}, O_{REF} = o_{REF}] \text{ max.}$$

which is equivalent to find $e \in \{e\}$ such that (using the Bayes theorem):

$$Pr[E = e]\ Pr[O_{FD} = o_{FD}, O_{REF} = o_{REF}/E = e] \text{ max.}$$

C. Energy function

The maximisation of this probability is equivalent to the minimisation of an energy function U which is the weighted sum of several terms [28]:

$$\begin{aligned} U(e, o_{FD}, o_{REF}) \;=\; & U_m(e) + \lambda_{FD} U_a(o_{FD}, e) + \\ & \lambda_{REF} U_a(o_{REF}, e) \end{aligned} \tag{1}$$

The model energy $U_m(e)$ may be seen as a regularization term that ensures spatio-temporal homogeneity of the masks of moving people and eliminates isolated points due to noise. Its expression resulting from the equivalence between MRF and Gibbs distribution is:

$$U_m(e) = \sum_{c \in C} V_c(e_s, e_r)$$

c denotes any of the binary cliques defined on the spatio-temporal neighbourhood of Fig. 4.

A binary clique $c = (s, r)$ is any pair of distinct sites in the neighbourhood, including the current pixel s and anyone of the neighbours r. C is the set of all cliques. $V_c(e_s, e_r)$ is an elementary potential function associated to each clique $c = (s, r)$. It takes the following values:

$$V_c(e_s, e_r) = \begin{cases} -\beta_r & \text{if } e_s = e_r \\ +\beta_r & \text{if } e_s \neq e_r \end{cases}$$

where the positive parameter β_r depends on the nature of the clique: $\beta_r = 20, \beta_r = 5, \beta_r = 50$ for spatial, past temporal

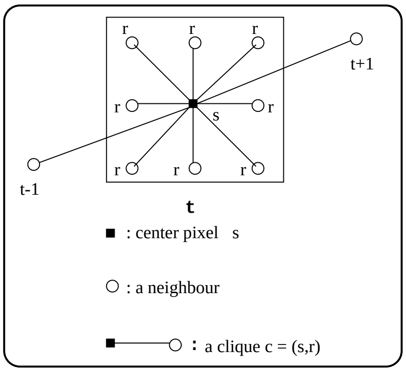


Fig. 4. Spatio-temporal neighbourhood and binary cliques.

and future temporal cliques respectively. Such values have been experimentally determined once and for all.

The link between labels and observations (generally noted O) is defined by the following equation:

$$o(s,t) = \Psi(e(s,t)) + n(s)$$

$$\text{where} \quad \Psi(e(s,t)) = \begin{cases} 0 & \text{if } e(s,t) = bg \\ \alpha > 0 & \text{if } e(s,t) = obj \end{cases}$$

and $n(s)$ is a Gaussian white noise with zero mean and variance σ^2. σ^2 is roughly estimated as the variance of each observation field, which is computed online for each frame of the sequence so that it is not an arbitrary parameter.

$\Psi(e(s,t))$ models each observation so that n represents the adequation noise: if the pixel s belongs to the static background, no temporal change occurs neither in the intensity function nor in the difference with the reference frame so each observation is quasi null; if the pixel s belongs to a moving person, a change occurs in both observations and each observation is supposed to be near a positive value α_{FD} and α_{REF} standing for the average value taken by each observation.

Adequation energies $U_a(o_{FD}/e)$ and $U_a(o_{REF}/e)$ are computed according to the following relations:

$$U_a(o_{FD}, e) = \frac{1}{2\sigma_{FD}^2} \sum_{s \in I} [o_{FD}(s,t) - \Psi(e(s,t))]^2$$

$$U_a(o_{REF}, e) = \frac{1}{2\sigma_{REF}^2} \sum_{s \in I} [o_{REF}(s,t) - \Psi(e(s,t))]^2$$

Two weighting coefficients λ_{FD} and λ_{REF} are introduced since the correct functioning of the algorithm results from a balance between all energy terms. $\lambda_{FD} = 1$ is set once and for all, this value does not depend on the processed sequence. λ_{REF} is fixed according to the following rule:

- $\lambda_{REF} = 0$ if $I_{REF}(s,t)$ does not exist: when no reference frame is available at pixel s, $o_{REF}(s,t)$ does not influence the relaxation process
- $\lambda_{REF} = 25$ if $I_{REF}(s,t)$ exists. This high value illustrates the confidence in the reference frame when it exists.

D. Relaxation

The deterministic relaxation algorithm ICM (Iterated Conditional Modes [29]) is used to find the minimum value of the energy function given by Equation (1). For each pixel in the image, its local energy is computed for each label (*obj* or *bg*). The label that yields a minimum value is assigned to this pixel. As the pixel processing order has an influence on the results, two scans of the image are performed in an ICM iteration, the first one from the top left to bottom right corner, the second one in the opposite direction. Since the greatest decrease of the energy function U occurs during the first iterations, we decide to stop after four ICM iterations. Moreover, one ICM iteration out of two is replaced by morphological closing and opening, see Fig. 3. It results in an increase of the processing rate without losing quality because the ICM process works directly on the observations (temporal frame differences) computed from the frame sequence and does not work on binarized observation fields. The ICM algorithm is iterative and does not insure the convergence towards the absolute minimum of the energy function, therefore an initialisation of the label field E is required: it results from a `logical or` between both binarized observation fields O_{FD} and O_{REF}. This initialisation helps converging towards the absolute minimum and requires two binarization thresholds which depend on the acquisition system and the environment type (indoor or outdoor).

Once this segmentation process is performed, the label field yields a segmentation mask for each video object present in the scene (single person or group of people). The segmentation masks are obtained through a connex component labelling of the segmented pixels whose label is *obj*. Fig. 5 shows an example of obtained segmentation in our system. The results are good, the person is not split and the boundaries are precise, even if there are some shadows around the feet.

(a) (b)

Fig. 5. Segmentation example. (a) original frame, (b) segmented frame.

For each video object, single person or group of people, once the segmentation mask is obtained, more low-level data are available and computed:

- surface: number of pixels of an object
- centre of gravity of the object
- SRBB: Segmentation Rectangular Bounding Box
- SPAB: Segmentation Principal Axes Box, whose directions are given by the principal axes of the object shape

After this first step of low-level information extraction, the next step after segmentation is basic temporal tracking.

III. BASIC TEMPORAL TRACKING

In many vision-based systems, it is necessary to detect and track moving people passing in front of a camera in real-time [1, 2]. Tracking is a crucial step in human motion analysis, for it temporally links features chosen to analyse and interpret human behaviour. Tracking can be performed for a single human or for a group, seen as an object formed of several humans or as a whole.

A. Our approach

The tracking method presented in this section is designed to be fast and simple. It is used mainly to help the face localisation step presented in the next section. Therefore it only needs to establish a temporal link between people detected at time t and people detected at time $t-1$. This tracking stage is based on the **computation of the overlap of the segmentation rectangular bounding boxes**. The segmentation rectangular bounding boxes are noted SRBBs. This method does not handle occlusions between people but allows the detection of temporal split and merge. In the case of a group of people, as there is only one video object composed of several people, this group is tracked as a whole in the same way as if the object was composed of a single person.

After the segmentation step, each SRBB should contain either a single person or several people, in the case of a merge. Only the general constraints of the system are assumed, in particular constraint n°2 (people are the only both **big** and **mobile objects**) and constraint n°3 (each person enters the scene **alone**).

As the acquisition rate of the camera is 30 fps we can suppose that the people in the scene have a small motion from one frame to the next, i.e. there is always a non-null overlap between the SRBB of a person at time t and the SRBB of this person at time $t-1$. Therefore a basic temporal tracking is possible by considering only the overlaps between detected boxes at time t and those detected at time $t-1$. We do not use motion compensation of the SRBBs because it would require motion estimation which is time-consuming.

In order to detect temporal split and merge and to ease the explanations, two types of objects are considered:

- SP: Single Person
- GP: Group of People

This approach is similar to the one used in [30] where the types: regions, people and group are used. When a new object is detected, with regard to constraint n°3 of the system, this object is assumed to be a SP human being. It is given a new ID (IDentification number). GP are detected when at least two SPs merge.

The basic temporal tracking between SRBBs detected on two consecutive frames (time $t-1$ and t) results from the combination of a forward tracking phase and a backward tracking phase. For the forward tracking phase, we look for the successor(s) of each object detected at time $t-1$ by computing the overlap surface between its SRBB and all the SRBBs detected at time t. In the case of multiple successors, they are sorted by decreasing overlap surface (the most probable successor is supposed to be the one with the greatest overlap surface). For the backward tracking phase, the procedure is similar: we look for the predecessor(s) of each object detected at time t. Considering a person P detected at time t: if P's most probable predecessor has P as most probable successor, a temporal link is established between both SRBBs (same ID). If not, we look in the sorted lists of predecessors and successors until a correspondence is found, which is always possible if P's box has at least one predecessor. If this is not the case, P is a new SP (new ID).

As long as an object, i.e. a single person or a group of people, is successfully tracked, without any temporal split or merge, its ID remains unchanged.

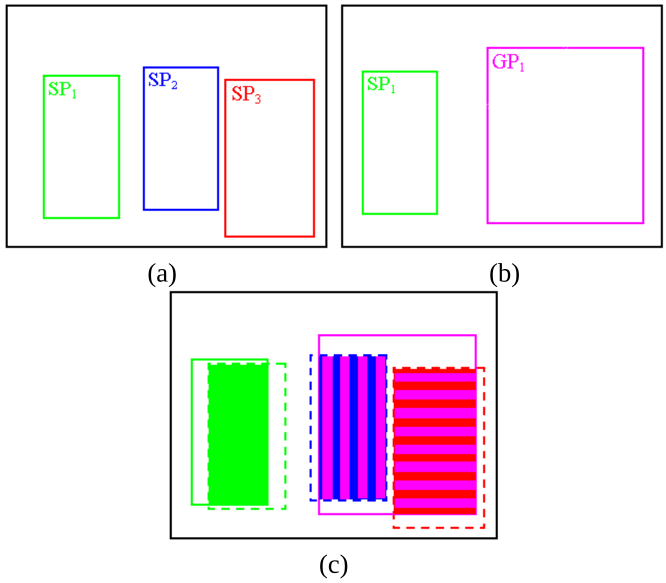


Fig. 6. Overlap computation. (a) frame at time $t-1$, (b) frame at time t and (c) overlap frame.

Fig. 6 illustrates the backward-forward tracking principle. On (a), three objects are segmented, all SP, and on (b), only two objects are segmented. On the overlap frame (c), the backward and forward tracking lead to a correct tracking for the object on the left side (there is only one successor and predecessor). It is tracked as a SP. For the object on the right side, the backward tracking yields two SP predecessors, and the forward tracking one successor. A merge is detected and it is a new group that will be tracked as a GP until it splits.

This basic temporal tracking is very fast and allows:

- **Segmentation problems correction**: If one SP has several successors, in case of a poor segmentation, we can merge them back into an SP and correct the segmentation.
- **GP split detection**: If a GP splits in several SPs, nothing is done, but a split is detected.
- **SP merge detection**: If several SPs merge, the resulting object has several SP predecessors so it is recognised as a GP and a merge is detected.

Fig. 7 shows frames of a video sequence where two people are crossing, when they are merging into a group and when this group is splitting. Segmentation results, SRBBs and trajectories of gravity centres are drawn on the original frames. The trajectories are drawn as long as there is no temporal split or merge, i.e. as long as the tracked object type does not change. In frame 124, tracking leads to SP P_1 on the left side and SP P_2 on the right side. In frame 125, a GP G_1, composed of P_1 and P_2, is detected. For the forward tracking phase between times 124 and 125, P_1 and P_2 have G_1 as only successor. For the backward tracking phase, G_1 has P_1 as first predecessor and P_2 as second predecessor. But, in this case, as P_1 and P_2 are SPs, a merge is detected. Therefore G_1 is a new GP, which will be tracked until it splits again. It is the opposite on frames 139 and 140. The GP G_1 splits into two new SPs P_3 and P_4 that are successfully tracked until the end.

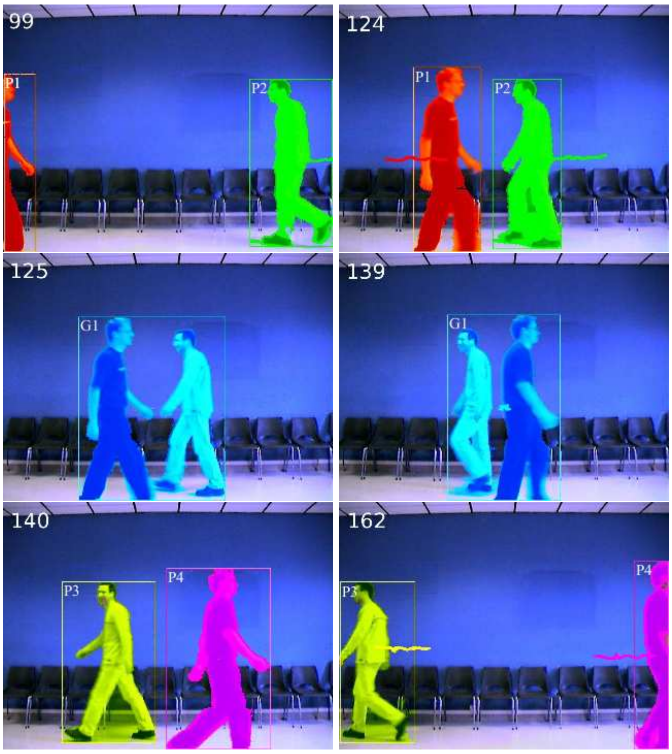


Fig. 7. Basic temporal tracking example. Frames 99, 124, 125, 139, 140 and 162 of two people crossing.

In the first tracking stage, a person may not be identified as a single entity from beginning to end if there are more than one people present in the scene. This will be done by the second tracking stage. The results of this processing step are the IDentification numbers (IDs), the object types (SP or GP), and the temporal split and merge information. Moreover, the trajectories for the successfully tracked objects are available.

In this paper, the presented results have been obtained after carrying out experiments on a great majority of sequences with one or two people, and on a few sequences with three. We consider that it is enough for the aimed applications (HCIs, indoor video surveillance and mixed reality applications). The constraint n°2 of the system specifies that people are the only both big and mobile objects in the scene. For this reason, up to three different people can be efficiently tracked with this basic temporal tracking method. If there are more than three people, it is difficult to determine, for instance, whether a group of four people have split into two groups of two people or into a group of three people and a single person.

After this basic temporal tracking processing step, the next step is face and hands localisation.

IV. FACE AND HANDS LOCALISATION

Numerous papers on human behaviour analysis focus on face tracking and facial features analysis [31, 32, 33]. Indeed, when looking at people and interacting with them, our gaze focuses on faces, as the face is our main expressive communication medium, followed by the hands and our global posture. Hand gesture analysis and recognition is also a large research field. The localisation of the face and of the hands, with right/left distinction, is also an interesting issue with respect to the considered applications. Several methods are available to detect faces [33, 34, 35]: using colour information [36, 37], facial features [38, 39], and also: templates, optic flow, contour analysis and a combination of these methods. It has been shown in those studies that skin colour is a strong cue for face detection and tracking and that it clusters in some well chosen colour spaces.

A. Our approach

With our constraints, for computing cost reasons, the same method has to be used to detect the face and the hands in order to achieve real-time processing. As features would be too complex to define for hands, a method based on colour is better suited to our application. When the background has a colour similar to the skin, this kind of method is perhaps less robust than a method based on body modelling. However, results have shown that the proposed method works on a wide range of backgrounds, providing efficient skin detection. In this paper, we present a robust and adaptive skin detection method working in the *YCbCr* colour space and based on an adaptive thresholding in the *CbCr* plane. Several colour spaces have been tested and the *YCbCr* colour space is one of those that yielded the best results [40, 41]. A method of selecting the face and hands among skin patches is also described. For this processing step, only the general constraints (n°1, 2 and 3) are assumed. When the static posture recognition processing step was developed, we had to define a reference posture (standing, both arms stretched horizontally), see section VI.*A*. Afterwards, we decided to use this reference posture, if it occurs and if necessary, to re-initialise the face and hands locations.

Fig. 8 summarises the face/hands localisation step.

B. Skin detection

This section describes the detection of skin pixels, based on colour information. For each SRBB (Segmentation Rectangular Bounding Box) provided by the segmentation step, we look for skin pixels. Only the segmented pixels inside the SRBBs are processed. Thanks to this, few background pixels (even if the background is skin colour-like) are processed.

A skin database is built, composed of the Von Luschan skin samples frame (see Fig. 9(a)) and of twenty skin frames (see examples Fig. 9(b)) coming from various skin colours hands or arms. The skin frames are acquired with the camera and frame grabber we use in order to take into account the white balance and the noise of the acquisition system.

Fig. 10 is a 2D plot of all pixels from the skin database on the *CbCr* plane with an average value of *Y*. It exhibits two lobes: the left one corresponds to the Von Luschan skin samples frame and the right one to the twenty skin samples acquired with our

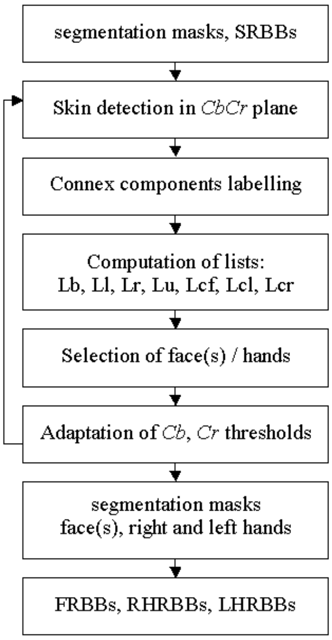
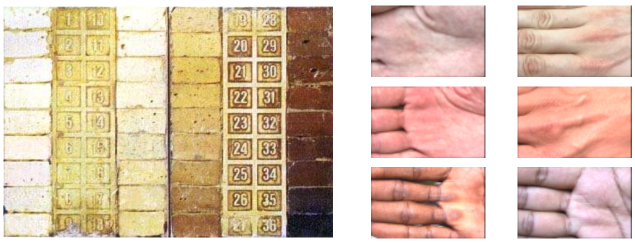


Fig. 8. Scheme of the face and hands localisation processing step.



(a) (b)

Fig. 9. Skin database. (a) Von Luschan frame, (b) 6 skin samples.

camera and frame grabber.

Fig. 11 shows an example of skin detection where optimal manually tuned thresholds were used. Results are good: face and hands (arms here) are correctly detected with accurate boundaries.

The *CbCr* plane is partitioned into two complementary areas: skin area and non-skin area. A rectangular model for the skin area shape yields a good detection quality with a low computing cost. It limits the required computations to a double thresholding (low and high) for each *Cb* and *Cr* component. As video sequences are acquired in the *YCbCr* 4:2:0 format, *Cb* and *Cr* components are sub-sampled by a factor of 2. The skin/non skin decision for a 4×4 pixels block of the segmented frame is taken after the computation of the average values of a 2×2 pixels block in each *Cb* or *Cr* sub-frame. Those mean values are then compared with the four thresholds. Computation is therefore even faster.

A rectangle containing most of our skin samples is defined by $Cb \in [86; 140]$ and $Cr \in [139; 175]$ (big rectangle of Fig. 10). This rectangle is centred on the mean values of the lobe corresponding to our skin samples frames to adjust the detection to our acquisition system. The right lobe is not completely included in the rectangle in order to avoid too much false detection. In [42] considered thresholds are slightly different ($Cb \in [77; 127]$ and $Cr \in [133; 173]$) which justifies the tuning of parameters to the first source of variability, i.e. the acquisition system and the lighting conditions. The second source of variability is the inter-individual skin colour. Each small rectangle of Fig. 10 only contains skin samples from a particular person in a given video sequence. Therefore it is also useful to automatically adapt the thresholds to each person during the detection process in order to improve the skin segmentation.

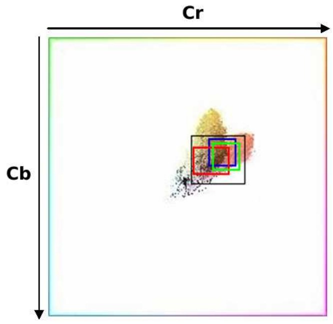
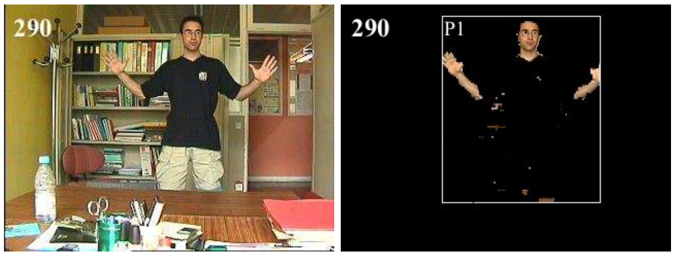


Fig. 10. 2D plot of all skin samples pixels.



(a) (b)

Fig. 11. Example of skin detection. (a) original frame, (b) skin detection.

Several papers detail the use of colour models, for instance Gaussian pdf in the *HSI* or *rgb* colour space [36] and perform an adaptation of model parameters. An evaluation of Gaussianity of *Cb* and *Cr* distributions was performed on the pixels of the skin database. As a result, approximately half of the distributions can not be reliably represented by a Gaussian distribution [41]. Therefore thresholds are directly adapted without considering any model.

Skin detection thresholds are initialised with (*Cb*,*Cr*) values defined by the big rectangle of Fig. 10. In order to adapt the skin detection to inter-individual variability, transformations of the initial rectangle are considered (they are applied separately to both dimensions *Cb* and *Cr*). These transformations are performed with respect to the mean values of the face skin pixels distribution of the considered person. Only the skin pixels of the face are used, as the face moves more slowly and is easier to detect than hands. This prevents the adaptation from being biased by detected noise or false hands detection. Three transformations are considered for the threshold adaptation:

- **Translation:** The rectangle is gradually translated towards the mean values of skin pixels belonging to the selected

face skin patch. The translation is of only one colour unit per frame in order to avoid transitions being too sharp. The translated rectangle is also constrained to remain inside the initial rectangle.
- **Reduction:** The rectangle is gradually reduced (also of one colour unit per frame). Either the low threshold is incremented or the high threshold is decremented so that the reduced rectangle is closer to the observed mean values of skin pixels belonging to the face skin patch. Reduction is not performed if the adapted rectangle reaches a minimum size (15×15 colour units).
- **Re-initialisation:** The adapted rectangle is reinitialised to the initial values if the adapted thresholds lead to no skin patch detection.

Those transformations are applied once to each detection interval for each frame of the sequence. As a result skin detection should improve over time. In most cases, the adaptation needs ~ 30 frames (~ 1 s of acquisition time) to reach a stable state.

C. Face and hands selection

This section proposes a method in order to select relevant skin patches (face and hands). Pixels detected as skin after the skin detection step are first labelled into connex components that can be either real skin patches or noise patches. All detected connex components inside a given SRBB are associated to it. Then, among these components, for each SRBB, skin patches (if present) have to be extracted from noise and selected as face or hands. To reach this goal several criteria are used. Detected connex components inside a given SRBB are sorted in decreasing order in lists according to each criterion. The left or right side of the lists are from the user's point of view.

Size and position criteria are:
- List of biggest components (Lb): face is generally the biggest skin patch followed by hands and other smaller patches are generally detection noise
- List of leftmost components (Ll): useful for left hand
- List of rightmost components (Lr): useful for right hand
- List of uppermost components (Lu): useful for face

Temporal tracking criteria are:
- List of closest components to last face position (Lcf)
- List of closest components to last left hand position (Lcl)
- List of closest components to last right hand position (Lcr)

Selection is guided by heuristics related to human morphology. For example, the heuristics used for the face selection are: the face is supposed to be the biggest, the uppermost skin patch and the closest to the previous face position. The face is the first skin patch to be searched for because it has a slower and steadier motion than both hands and therefore can be found more reliably than hands. Then the skin patch selected as the face is not considered any longer. After the face selection, if one hand was not found in the previous frame, we look for the other first. In other cases hands are searched without any *a priori* order.

Selection of the face involves (Lb, Lu, Lcf), selection of the left hand involves (Lb, Ll, Lcl) and selection of the right hand involves (Lb, Lr, Lcr). The lists are weighted depending on the skin patch to find and if a previous skin patch position exists. The list of biggest components is given a unit weight. All other lists are weighted relatively to this unit weight. If a previous skin patch position exists, the respective list of closest components is given a triple weight. As the hand does not change side from one frame to another, if the skin patch previous position is on the same side as the respective side list (Lr for the right hand), this list is given a double weight. The top elements of each list are considered as likely candidates. When the same element is not at the top of all lists, the next elements in the list(s) are considered. The skin patch with the **maximum weighted lists rank sum** is finally selected.

For the face, in many cases there is a connex component that is at the top of those three lists. In the other cases, Lcf (tracking information) is given the biggest weight because face motion is slow and steady. The maximum rank considered in other lists is limited to three in order to avoid unlikely situations and poor selection.

After selection, the face, right and left hands rectangular bounding boxes are also computed (noted respectively FRBB, RHRBB and LHRBB). For the face skin patch, considering its slow motion, we add the constraint of a non-null rectangular bounding box overlap with its successor. This helps to handle situations where a hand passes in front of the face. Moreover, if the person is in the reference posture (see section VI), this posture is used to correctly re-initialise the locations of the face and of the hands in the case of a poor selection or a tracking failure.

Fig. 12 illustrates some results of face/hands localisation. Skin detection is performed inside the SRBB. Face and hands are correctly selected and tracked as shown by the small rectangular bounding boxes. Moreover, even if the person crosses his arms (frames 365 and 410), the selection is still correct.

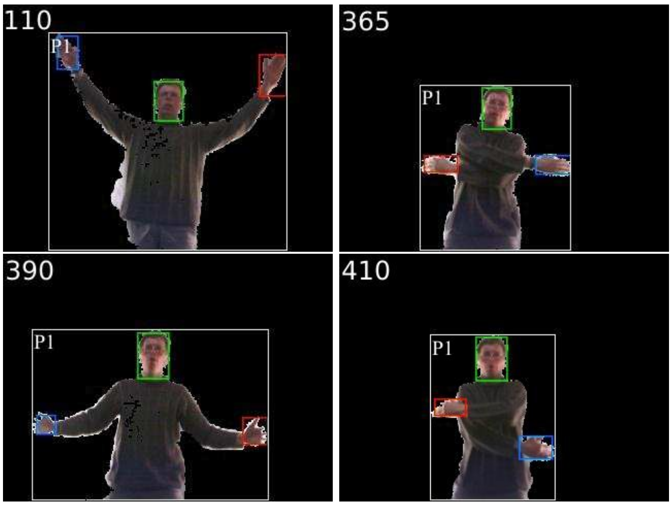


Fig. 12. Face and hands localisation. Frames number 110, 365, 390 and 410.

For each object in the scene, the low-level data available at the end of this processing step are the three selected skin patches segmentation masks (face, right hand and left hand) and their rectangular bounding boxes (noted respectively FRBB, RHRBB and LHRBB). In the next section, an advanced tracking dealing with occlusions problem is presented thanks to the use of face-related data. The data about hands are not used in the rest of this paper but have been used in other applications, like the *art.live* project [3].

V. KALMAN FILTERING-BASED TRACKING

The basic temporal tracking presented in section III does not handle temporal split and merge of people or groups of people. When two tracked people merge into a group, the basic temporal tracking detects the merge but tracks the resulting group as a whole until it splits. Then people in the group are tracked again but without any temporal link with the previous tracking of individuals. In Fig. 7 two people P_1 and P_2 merge into a group G_1. When this group splits again in two people, they are tracked as P_3 and P_4, not as P_1 and P_2. Temporal merge and occlusion make the task of tracking and distinguishing people within a group more difficult [30, 43, 44]. This section proposes an overall tracking method which uses the combination of partial Kalman filtering and face pursuit to track multiple people in real-time even in case of complete occlusions [45].

A. Our approach

We present a method that allows the tracking of multiple people in real-time even when occluded or wearing similar clothes. Apart from the general constraints of the system (n°1, 2 and 3), no other particular hypothesis is assumed here. We do not segment the people during occlusion but we obtain bounding boxes estimating their positions. This method is based on partial Kalman filtering and face pursuit. The Kalman filter is a well-known optimal and recursive signal processing algorithm for parameters estimation [46]. With respect to a given model of parameters evolution, it computes the predictions and adds the information coming from the measurements in an optimal way to produce *a posteriori* estimation of the parameters. We use a Kalman filter for each new detected person. The global motion of a person is supposed to be the same as the motion of this person's face. Associated with a constant speed evolution model, this leads to a state vector $\underline{x}$ of ten components for each Kalman filter: the rectangular bounding boxes of the person and of his/her face (four coordinates each) and two components for the 2D apparent face speed:

$$\underline{x}^T = (x_{pl}, x_{pr}, y_{pt}, y_{pb}, x_{fl}, x_{fr}, y_{ft}, y_{fb}, v_x, v_y).$$

In $\underline{x}^T$ expression, p and f respectively stand for the person and face rectangular bounding box, l, r, t and b respectively stand for left, right, top and bottom coordinate of a box. v_x and v_y are the two components for the 2D apparent face speed. The evolution model leads to the following Kalman filter evolution matrix:

$$A_t = A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Fig. 13 summarises the Kalman filtering-based tracking processing step.

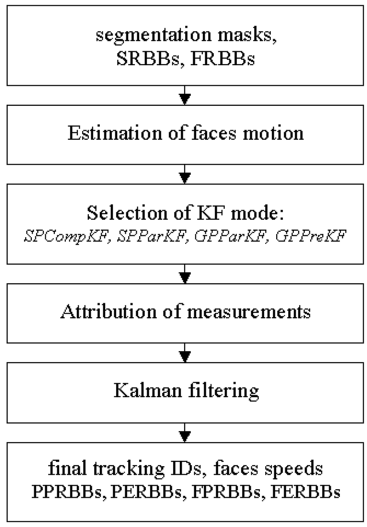


Fig. 13. Scheme of the Kalman filtering-based tracking processing step.

B. Face motion estimation

For each face that is detected, selected and located at time $t-1$ by the method presented in section IV, we estimate a face motion from $t-1$ to t by block-matching in order to obtain the 2D apparent face speed components v_x and v_y. For each face, the pixels inside the FRBB (Face Rectangular Bounding Box) are used as the estimation support.

C. Notations

The segmentation step may provide SRBBs (Segmentation Rectangular Bounding Boxes) that can contain one or several people in it (in the case of a merge) whereas the Kalman state vector (and therefore the Kalman person rectangular bounding box) is defined for a single person. Therefore three different person rectangular bounding boxes exist and are associated to each person:

- one Segmentation Rectangular Bounding Box (SRBB) provided by the segmentation step,
- one Person Predicted Rectangular Bounding Box (PPRBB) predicted by Kalman filtering and
- one Person *a posteriori* Estimated Rectangular Bounding Box (PERBB) estimated by Kalman filtering.

In a similar way, three different face rectangular bounding boxes exist and are associated to each person:

- one Face Rectangular Bounding Box (FRBB) provided by the face localisation step,
- one Face Predicted Rectangular Bounding Box (FPRBB) predicted by Kalman filtering and
- one Face *a posteriori* Estimated Rectangular Bounding Box (FERBB) estimated by Kalman filtering.

D. Kalman filtering modes

Measurements that are injected into the Kalman filter come from the SRBBs, the FRBBs and the face motion estimations. All the measurements are not necessarily available. For instance, if two people have just merged into a group, some measurements are not available, on the group SRBB, for each person's PPRBB estimation (for example, one side measurement will not be available).

Depending on the objects types and available measurements, there are four Kalman filtering modes:

1. *SPCompKF*: Single Person Complete Kalman Filtering
2. *SPParKF*: Single Person Partial Kalman Filtering
3. *GPParKF*: Group of People Partial Kalman Filtering
4. *GPPreKF*: Group of People Predictive Kalman Filtering

First, we must determine if we are in a single person mode or a group of people mode, i.e. if the person SRBB contains only one person or not. This is given by the basic temporal tracking step, as we can detect a merge between two SP objects, we know if there is one person or more in each SRBB.

If the SRBB contains only one person, all measurements used for the PPRBB estimation are available. Then either the face was correctly located at times $t-1$ and at time t or not. If so, we are in *SPCompKF* mode as all state vector measurements are available. Otherwise we are in *SPParKF* mode as some face-related measurements are not available.

If the SRBB contains several people, some measurements are not available for the PPRBBs estimation. Depending on whether there is only one face overlapped by the PERBB or not, we are respectively in *GPParKF* mode or in *GPPreKF* mode.

D.1 Single Person Complete Kalman Filtering mode

This mode is selected when there is no temporal merge and all face-related measurements are available:

- The SRBB contains only one person (all measurements for the PPRBB estimation are available)
- The person's face is located at time t (all measurements for the FPRBB estimation are available)
- The person's face has been located at time $t-1$ (face speed estimation measurements are available)

In this mode, the Kalman filtering is carried out for all state vector components.

D.2 Single Person Partial Kalman Filtering mode

This mode is selected when there are no temporal merge but some or all face-related measurements are not available. If so, face localisation step has failed at time $t-1$ and/or at time t, leading to unavailable measurements.

When there are unavailable measurements, two choices are possible. The first is to perform a Kalman filtering only on the available measurements and the other is to replace the unavailable measurements. Performing a Kalman filtering only on available measurements is a difficult issue for code implementation, as all matrix sizes have to be predicted in order to take into account all possible cases. Replacing unavailable measurements by predictions is a simple and intuitive way of performing a Kalman filtering when observations (available measurements) are missing. Hence, in order to perform a Kalman filtering for all state vector components in one computation, when there are unavailable measurements, they are replaced by predictions. Doing so does not seem to greatly influence the results because the variances of estimation errors are only of a few pixels, with respect to available measurements.

In this mode, the filtering is carried out for all components, including those that have been replaced by predicted values.

D.3 Group of People Partial Kalman Filtering mode

This mode is selected when there are temporal merge(s) (i.e. some measurements are not available for the PPRBB estimation) and when the PERBB overlaps a unique face.

As the SRBB contains a group of people, available measurements can be used for different PPRBBs. The attribution of available measurements to one person in a group is performed in two steps by comparing the group SRBB and each person PPRBB centres and sides coordinates. The principle of measurements attribution is illustrated on frame 203 of Fig. 14.

In the first step, we compare the coordinates of the PPRBBs centres to the coordinates of the SRBB centre. With respect to the SRBB quarter where each PPRBB centre is located, the two closest sides coordinates are used as measurements for the corresponding PPRBB estimation. For example, on frame 203 of Fig. 14, if two people have just merged (hands touching), we have only four measurements available (instead of eight) that can be used as observations for the two PPRBBs. With the first step, the person P_1 will have the left and bottom sides coordinates as measurements, the person P_2 will have the right and bottom sides coordinates. Thanks to this step, we are sure that at least two measurements are used for each PPRBB estimation.

In the second step, we compare each PPRBB side coordinate to the corresponding SRBB side. If the distance between both is smaller than a threshold, depending on each PPRBB surface, and if it has not already been taken into account, the corresponding SRBB side coordinate is added to the measurements used for the PPRBB estimation. With this step, in our example, the person P_1 receives the top side coordinate of the SRBB as an added measurement. This step generally allows adding one or two measurements in order to perform a better estimation.

In the example of Fig. 14, the left, top and bottom side measurements of the SRBB will be used as measurements for the PPRBB on the left side (person P_1). The right and bottom side measurements will be used as measurements for the PPRBB on the right side (person P_2). As for the bottom side measurement in the example, some measurements can be used for different people. For each person, in this *GPParKF* mode, we generally have two or three available measurements (up and/or down side(s) and one side measurements).

If some face-related measurements are unavailable, Kalman predicted values replace the missing measurements. The filtering is performed as long as the PERBB contains a unique face. If the PERBB overlaps more than one face, even partially, the Kalman filter works in *GPPreKF* mode since the face localisation step could provide wrong positions.

D.4 Group of People Predictive mode

This mode is selected when temporal merge(s) occur (i.e. some measurements are not available for the PPRBB estimation) and when the PERBB overlaps more than one face.

No measurements are taken into account. All the state vector components are predicted according to the last face speed estimation, i.e. only the Kalman filter predictions equations are used. The Kalman filter works in *GPPreKF* mode until a unique face is again overlapped by one of the PERBBs, leading back to the *GPParKF* mode.

E. Results

Fig. 14 illustrates a successful multiple people tracking performed on a video sequence in which two people are crossing and turning one around the other. In this sequence, the 2D apparent directions and speeds are not constant and, at some moments, a person is completely occluded, see for instance frame 212. Segmented and tracked people are visible on the original frames of the sequence. SP or GP SRBBs are drawn in white lines, PERBBs and FERBBs in dashed lines. Frames 200 and 228 show a *SPCompKF* mode tracking with all measurements available for the Kalman filters before the merge (frame 203) and after the split (frame 228). Frames 212 and 219 illustrate the tracking in a *GPPreKF* mode when one face is occluded. Frames 203 and 221 (just before the split) illustrate the tracking in *GPParKF* mode.

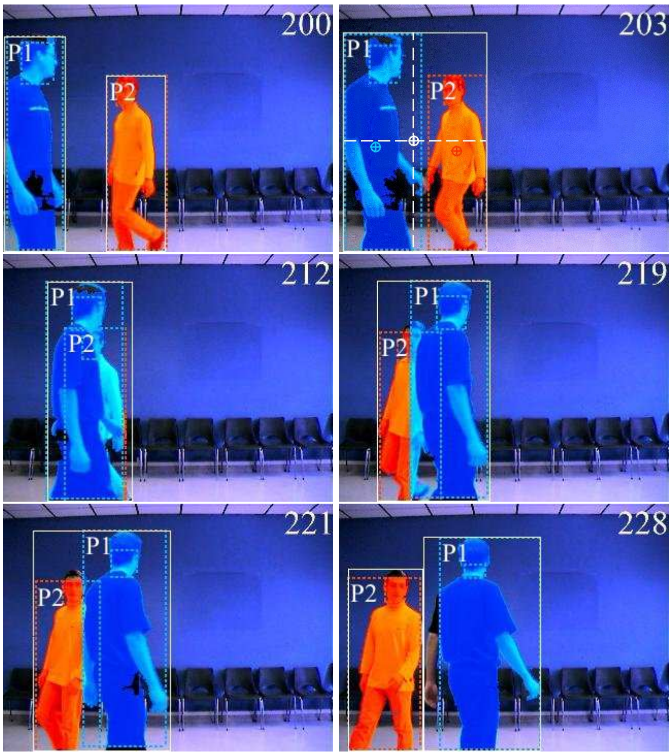


Fig. 14. Example of multiple people tracking with complete occlusion.

In single person Kalman filtering modes, *SPCompKF* mode and *SPParKF* mode, the person final tracking ID is the same as the basic temporal tracking ID, because there are no temporal split or merge. In group of people modes, *GPParKF* and *GPPreKF*, the final tracking IDs are not updated with the basic temporal tracking IDs, as temporal split and merge yield new IDs. Therefore it is possible to track multiple people even under complete occlusions. The extracted information for this processing step consists of the final tracking IDs, the face speed estimation, the PPRBBs, the PERBBs, the FPRBBs and the FERBBs, i.e. the predicted and *a posteriori* estimated rectangular boxes of the person and of his/her face.

This section presented the last processing step for low-level data extraction. Part of the data will now be used for higher-level processing.

VI. High-level human behaviour interpretation: Static posture recognition

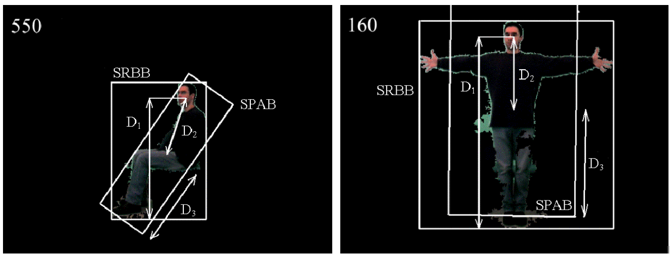
After having successfully tracked people, the problem of understanding human behaviour follows naturally. It involves action/pose recognition and description. The three main approaches used for human behaviour analysis used are Dynamic Time Warping (DTW) [47], Hidden Markov Models (HMMs) [48] and Neural Networks (NNs) [49]. Most of the research work done on the human body as a whole is mainly gait analysis and recognition, or recognition of simple interactions between people, or between people and objects. In this section, we present a method to recognise a set of four static human body postures (standing, sitting, squatting and lying) thanks to data fusion using the belief theory [50, 51].

The belief theory has been used for facial expression classification ([52, 53]) but not for posture recognition in human motion analysis. The TBM (Transferable Belief Model) was introduced by Smets in [54, 55]. It follows the works of Dempster [56] and Shafer [57]. The main advantage of the belief theory is the possibility to model data imprecision and conflict (a conflict occurs when measurements used for recognition yield contradictory results). It is also not computationally expensive, compared to HMMs and, as doubt (the possibility of recognising a union of postures instead of a unique one) is taken into account, leads to a low false alarm rate.

A. Our approach

Static recognition is based on information obtained by dynamic sequence analysis. For this processing step, we assume the general constraints of the system (n°1, 2 and 3) and also two more hypotheses:

- Each person has to be at least once in a **reference posture**, standing with both arms stretched horizontally, also known as the “Da Vinci Vitruvian Man posture”, see Fig. 15b.
- Each person is to be filmed **entirely** (**not occluded**).



(a) (b)

Fig. 15. Examples of distances D_i. (a) sitting posture, (b) reference posture.

Three distances are computed, see Fig. 15: D_1 the vertical distance from the FRBB centre to the SRBB bottom, D_2 the distance from the FRBB centre to the SPAB centre (gravity centre) and D_3 the SPAB semi great axe length. Each distance D_i is normalised with respect to the corresponding distance D_i^{ref} obtained when the person is observed in the reference posture in order to take into account the inter-individual variations of height and the distance of the person with respect to the camera. The measurements are noted $r_i = D_i/D_i^{ref}$ $(i = 1 \ldots 3)$.

B. Belief theory

The belief theory approach needs the definition of a world Ω composed of N disjunctive hypotheses H_i. Here the hypotheses are the following four static postures: standing (H_1), sitting (H_2), squatting (H_3), and lying (H_4). If the hypotheses are exhaustive, Ω is a closed world, i.e. the truth is necessarily in Ω. In this paper, we consider an open world, as all possible human body postures can not be classified in the considered postures. We add a hypothesis for the unknown posture class (H_0), but this hypothesis is not included in Ω. H_0 is a reject class: if we cannot recognise a posture between our considered postures, we will recognise an unknown posture. Therefore we have Ω={H_1, H_2, H_3, H_4} and H_0. In this theory, we consider the 2^N subsets A of Ω. In order to express the confidence degree in each subset A without favouring one of its composing elements, an elementary belief mass $m(A)$ is associated to it.

The m function, or belief mass distribution, is defined by:

$$\begin{array}{llll} m: & 2^{\Omega} & \longrightarrow & [0;1] \\ & A & \longmapsto & m(A) \quad \text{with} \quad \sum_{A\in 2^{\Omega}} m(A) = 1 \end{array}$$

B.1 Modelling

A model has to be defined for each measurement r_i in order to associate an elementary belief mass to each subset A, depending on the value of r_i. In a similar way to what was proposed in [52], two different model types are used (see Fig. 16). The first model type is used for r_1 and the second for r_2 and r_3.

The first model type is based on the idea that the lower the face of a person is located, the closer the person is to the lying posture. Conversely, the higher the face is located, the closer the person is to the standing posture. Depending on the value of r_1, either a single posture is recognised or the combination of a single posture and a union of two postures. In this last case the respective zones illustrate the imprecision and the uncertainty of the models. For example (see Fig. 16a):

r_1 value	H_i recognised	non-null belief masses
$f < r_1$	H_1	$m_{r_1}(H_1) = 1$
$\frac{e+f}{2} < r_1 < f$	$H_1, H_1 \cup H_2$	$m_{r_1}(H_1) + m_{r_1}(H_1 \cup H_2) = 1$
$e < r_1 < \frac{e+f}{2}$	$H_1 \cup H_2, H_2$	$m_{r_1}(H_1 \cup H_2) + m_{r_1}(H_2) = 1$
etc.	etc.	etc.

The second model type is based on the idea that squatting is a compact human shape, whereas sitting is a more elongated shape. Standing and lying are even more elongated shapes. The thresholds $g - j$ are different for r_2 and r_3. Depending on the value of each measurement r_2 or r_3, the system can set non-null belief masses to the single posture H_3, to the union of all postures (Ω corresponds to $H_1 \cup H_2 \cup H_3 \cup H_4$ here), to the subset standing, sitting or lying ($H_1 \cup H_2 \cup H_4$) or to two of the previous subsets.

B.2 Data fusion

The aim is to obtain a belief mass distribution $m_{r_{123}}$ that takes into account all available information (the belief mass distribution of each r_i). It is computed by using the conjunctive combination rule called **orthogonal sum** proposed by Dempster [56].

The orthogonal sum $m_{r_{ij}}$ of two distributions m_{r_i} and m_{r_j} is defined, for each A subset of 2^{Ω}, as follows:

$$m_{r_{ij}} = m_{r_i} \oplus m_{r_j} \qquad (2)$$

$$m_{r_{ij}}(A) = \sum_{B\in 2^{\Omega}, C\in 2^{\Omega}, B\cap C = A} m_{r_i}(B).m_{r_j}(C) \qquad (3)$$

The orthogonal sum is associative and commutative, so the order of the belief mass distributions fusion does not matter.

In case when $m_{r_{123}}(\emptyset) \neq 0$, $\emptyset$ being the empty set, there is a **conflict**, which means that the chosen models give contradictory results. This usually happens when some of the r_i are in the transition zones of the models. With these models, the subset with the maximum number of elements that can be obtained at the end of the data fusion process is a union of two postures. Therefore, subsets with three elements or Ω itself can not be obtained after fusion. Hence, we are sure that, in the worst case, there will be a possible confusion between two postures and not more. This is compliant with respect to the considered postures: it is difficult to imagine, for example, that a person can be simultaneously either standing, sitting or lying.

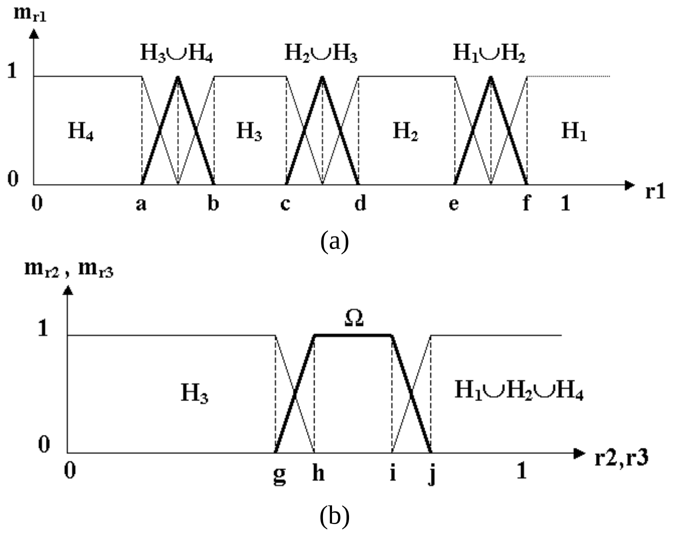


Fig. 16. Belief models (a) first model used for m_{r_1}, (b) second model used for m_{r_2} and m_{r_3}. H_i defines recognised posture(s).

B.3 Decision

The decision is the final step of the process. Once all the belief mass distributions have been combined into a single one, here $m_{r_{123}}$, there is a choice to make between the different hypotheses H_i and their possible combinations. A criterion defined on the final belief mass distribution is generally optimised to choose the classification result $\hat{A}$. For example, if the criterion is the belief mass $\hat{A} = \arg\max_{A\in 2^{\Omega}} m_{r_{123}}(A)$. Note that $\hat{A}$ may not be a singleton but a union of several hypotheses or even the empty set. In this paper, the hypothesis H_0 is chosen if the classification result is the empty set $\emptyset$, i.e. $m_{r_{123}}(\emptyset)$ is maximum. There are other criteria used to make a decision: the belief, the plausibility etc. [54].

C. *Posture recognition results*

In order to evaluate the static posture recognition performances, two sets of video sequences are used, a training set and a test set. The training set consists of 12 different video sequences representing $\sim$5000 frames. 6 different people are filmed twice in the same 10 successive postures. People are of various heights, between $1.55\ m$ and $1.95\ m$, in order to take into account the variability of heights and improve the robustness. The constraints are to be in "natural" postures in front of the camera. The statistics (means μ and standard deviations σ) of the three measurements r_i are computed over the training set to find the thresholds (see Fig. 16) that yield a minimum of conflict. These most suitable thresholds are defined by the comparison of the $\mu \pm 2\sigma$ computed for the respective postures or set of postures. This expertise step was performed by a human operator. In fact, one of the hardest steps in the belief theory is to find models (or thresholds) that lead to a minimum of conflicts. The test set consists of 12 other video sequences representing $\sim$11000 frames. 6 other people, also of various heights, are filmed twice in different successive postures. In order to test the limits of the system, people are allowed to move the arms, sit sideways and even be in postures that do not often occur in everyday life, for instance squatting with arms raised above the head. Results are computed on frames of the video sequences where the global body posture is static, i.e. the person's torso and legs are approximately still. We present the classification results obtained when using the **maximum belief mass** as criterion. Comparison between criteria and subsequent classifiers is available in [51]. Training step and test step recognition rates are available in Tables I and II. Columns show the real posture and lines the postures recognised by the system.

TABLE I
TRAINING STEP CONFUSION MATRIX

System$\backslash H$	H_1	H_2	H_3	H_4
H_0	0%	0.1%	0%	0%
H_1	**100%**	0%	0%	0%
$H_1 \cup H_2$	0%	0%	0%	0%
H_2	0%	**95.9%**	1.0%	0%
$H_2 \cup H_3$	0%	2.1%	4.0%	0%
H_3	0%	1.9%	**95.0%**	0%
$H_3 \cup H_4$	0%	0%	0%	0%
H_4	0%	0%	0%	**100%**

Training step: As the thresholds of the belief models are generated from the r_i statistical characteristics computed over the same set of video sequences, the results are very good. The average recognition rate is **97.7%**. There is only 0.1% of occurring conflicts on more than 5000 frames. There are no problems recognising the standing or the lying postures. The sitting and the squatting postures are also well recognised even if there is a little doubt between both.

Test step: There are more recognition errors but the results show a good global recognition rate. The average recognition rate is **78.1%**. There are never any problems recognising the standing or the lying postures. For the sitting and the squatting postures, there are more errors, especially when people have their arm(s) raised over their head or sit sideways. The reasons are that these postures are quite alike and that not everybody sits and/or squats in the same way, hands on knees or touching ground, back bent or straight etc. These facts yields more conflicts, near 15%. There are also more postures that lead to the doubt $H_2 \cup H_3$. Nevertheless, the recognition rates are very close between H_2 vs H_2 and H_3 vs H_3.

TABLE II
TEST STEP CONFUSION MATRIX

System$\backslash H$	H_1	H_2	H_3	H_4
H_0	0%	10.3%	5.0%	0%
H_1	**99.5%**	0.4%	0%	0%
$H_1 \cup H_2$	0.5%	0%	0%	0%
H_2	0%	**56.3%**	20.3%	0%
$H_2 \cup H_3$	0%	27.1%	18.0%	0%
H_3	0%	5.9%	**56.7%**	0%
$H_3 \cup H_4$	0%	0%	0%	0%
H_4	0%	0%	0%	**100%**

Fig. 17 illustrates some results of various static postures recognition. The SRBB, the SPAB, the FRBB and the D_2 distance are drawn in white on the segmented frame.

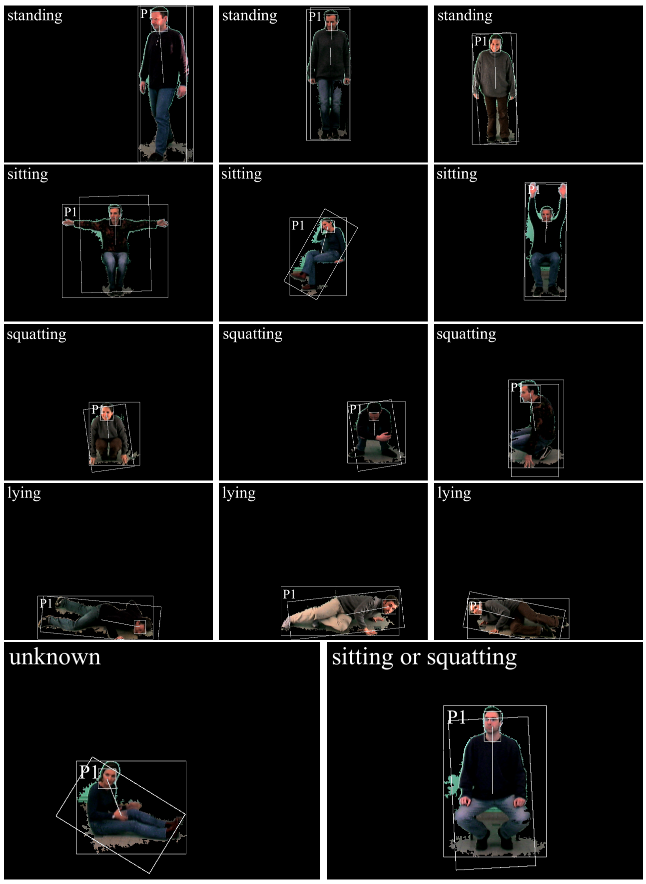


Fig. 17. Examples of static posture recognition.

VII. Conclusion, discussion and perspectives

A. Conclusion

We have presented in this paper a real-time system for multiple people body analysis and behaviour interpretation. The processing rate of the whole system, obtained on a PC running at 3.2 GHz, is ∼26 fps for 640×480 resolution (∼65 fps for 320×240). Compared with other similar systems like W^4 [1] and Pfinder [2], that surely meet the requirements to perform a similar task, our system proposes relatively different approaches for dealing with the various processing steps and their inherent problems. It is generic enough to be used for several types of applications in either indoor or outdoor environment. For outdoor environments, some of the algorithms would need to be improved, with regard to the problems that can arise when acquisition conditions greatly vary. As long as the people are not too numerous and remain the main objects, the results should be fairly reliable.

This system can be used for mixed reality applications with perceptual human-computer interfaces. In front of a single static camera, in an indoor environment, a single person or several people can interact with a virtual environment and control it by their movements. The proposed system for mixing real and virtual worlds by image processing without invasive systems as markers etc. yields results with a suitable precision. It is fast enough for a responsive system that includes human-computer interaction and is relatively user-friendly. The other possible application is the monitoring of elderly people at home or in hospital rooms. One could detect for instance that someone has fallen down or has been sitting for too long. Considering elderly people, their postures should be similar to the training set ones of the static posture recognition step. In these conditions, the system should be reliable enough to succeed in this monitoring as the training recognition rates are very good. Nevertheless, tests must still be performed and implemented source code improved.

B. Discussion and perspectives

The main advantages of the 2D segmentation step is that it yields smooth and regular segmentation masks and that the reference frame can be built even if the scene is not empty at the beginning. For indoor applications, a reference frame can be easily acquired when there is nobody present in the scene. No particular shadow processing is performed but some shadow models based on colour with invariant techniques could be used [58].

The tracking step, composed of the basic temporal tracking and of the Kalman filtering-based tracking, is very fast and handles partial or even complete occlusion problems. The tracking should still be efficient if people were occluded by fixed objects, as long as their global motion remains coherent with their face motion. If the people change direction or speed during the occlusion, the tracking results depend on the duration of the occlusion and on the other people's motion. In Fig. 14, the two people are turning one around the other and the tracking succeeds for this non-constant moving directions and speeds.

Using an adaptive thresholding in the *YCbCr* colour space, the skin detection process is robust enough to provide very good results even on complex or skin colour-like backgrounds. Hence localisation is generally accurate. It is fast and distinguishes the right vs left hand. Skin models are generally sensitive to the acquisition system and lighting conditions (output colour space, white balance and noise of the camera etc.). The presented thresholds have been tested in different indoor environments and performed reliably. Nevertheless, tuning them with respect to another given system (other camera, outdoor environment etc.) can yield better results. Results accuracy can be degraded when worn clothes are close to skin colours.

The higher-level interpretation step, static posture recognition, has also shown good recognition results. The approach we use is similar to a method based on shapes, because we consider the elongation and the compactness of the person's shape. Nevertheless, no explicit comparison has been performed. The main limitation is that, if the distance to the camera changes significantly, the person may have to perform again the reference posture. Using a stereo camera could solve this problem and avoid assuming the hypothesis of not being occluded.

Among the perspectives of this work, there is dynamic posture recognition. We plan to enhance the method by adding a dynamic analysis of the measurements temporal evolution. Concerning the analysis of human body parts, the feet positions could be computed after segmentation using geodesic distance maps [59]. Currently under development, there is an avatar control application with the real-time animation of a skeleton using the face and hands positions and the recognised posture. Work on gaze direction and facial expressions analysis is also under development [53, 60]. A long-term perspective is the fusion of multiple media with several cameras and microphones. This could lead to advanced perceptual human-computer interfaces and a lot of subsequent applications.

References

[1] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: who? when? where? what? a real time system for detecting and tracking people," in *IEEE-C-AFGR*, April 1998, pp. 222–227.

[2] C. R. Wren, A. Azarbayejani, T. J. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE-T-PAMI*, vol. 19, no. 7, pp. 780–785, July 1997.

[3] Website of the *art.live* project: IST Project 10942, "Architecture and authoring tools prototype for living images and new video experiments, http://www.transfiction.net/artlive/," 2002.

[4] Website of SIMILAR Network of excellence: the European taskforce creating human-machine interfaces similar to human-human communication, "http://www.similar.cc/," 2003.

[5] K. Aizawa and T. S. Huang, "Model-based image-coding: Advanced video coding techniques for very-low bit-rate applications," *PIEEE*, vol. 83, no. 2, pp. 259–271, February 1995.

[6] N. D. Doulamis, A. D. Doulamis, and S. D. Kollias, "Efficient content-based retrieval of humans from video databases," in *International Workshop on RATFG*, September 1999, pp. 89–95.

[7] N. Gehrig, V. Lepetit, and P. Fua, "Golf club visual tracking for enhanced swing analysis," in *British Machine Vision Conference*, September 2003.

[8] M. Kohle, D. Merkl, and J. Kastner, "Clinical gait analysis by neural networks: issues and experiences," in *IEEE 10th Symposium on Computer-Based Medical Systems*, 1997, pp. 138–143.

[9] P. Maes, T. J. Darrell, B. Blumberg, and A. P. Pentland, "The alive system: Wireless, full-body interaction with autonomous agents," *ACM Multimedia Sytems*, vol. 5, no. 2, pp. 105–112, March 1997.

[10] C. R. Wren, F. Sparacino, A. J. Azarbayejani, T. J. Darrell, T. E. Starner, A. Kotani, C. M. Chao, M. Hlavac, K. B. Russell, and A. P. Pentland, "Perceptive spaces for performance and entertainment: Untethered interaction using computer vision and audition," *Applied Artificial Intelligence*, vol. 11, no. 4, pp. 267–284, June 1997.

[11] D. M. Gavrila, "The visual analysis of human movement: a survey," *CVIU*, vol. 73, no. 1, pp. 82–98, January 1999.

[12] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *CVIU*, vol. 73, no. 3, pp. 428–440, March 1999.
[13] A. Pentland, "Looking at people: sensing for ubiquitous and wearable computing," *IEEE-T-PAMI*, vol. 22, no. 1, pp. 107–119, January 2000.
[14] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *CVIU*, vol. 81, no. 3, pp. 231–268, March 2001.
[15] L. Wang, W. M. Hu, and T. N. Tan, "Recent developments in human motion analysis," *PR*, vol. 36, no. 3, pp. 585–601, March 2003.
[16] J. J. Wang and S. Singh, "Video analysis of human dynamics: a survey," *Real-Time Imaging*, vol. 9, no. 5, pp. 321–346, October 2003.
[17] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa, "A system for video surveillance and monitoring," *CMU-RI-TR-00-12*, May 2000.
[18] V. Nair and J. J. Clark, "Automated visual surveillance using hidden markov models," in *Vision Interface*, 2002, pp. 88–94.
[19] A. Mitiche and P. Bouthemy, "Computation and analysis of image motion: a synopsis of current problems and methods," *IJCV*, vol. 19, no. 1, pp. 29–55, July 1996.
[20] H. H. Nagel, "Formation of an object concept by analysis of systematic time variations in the optically perceptible environment," *CGIP*, vol. 7, no. 2, pp. 149–194, April 1978.
[21] T. Aach, A. Kaup, and R. Mester, "Statistical model-based detection in moving videos," *Signal Processing*, vol. 31, no. 2, pp. 165–180, 1993.
[22] P. Sangi, J. Heikkila, and O. Silven, "Motion analysis using frame differences with spatial gradient measures," in *IEEE-C-ICPR*, August 2004, vol. 4, pp. 733–736.
[23] W. Long and Y. H. Yang, "Stationary background generation: An alternative to the difference of two images," *PR*, vol. 23, no. 12, pp. 1351–1359, 1990.
[24] M. Seki, T. Wada, H. Fujiwara, and K. Sumi, "Background subtraction based on cooccurrence of image variations," in *IEEE-C-CVPR*, June 2003, vol. 2, pp. 65–72.
[25] D. S. Lee, "Effective gaussian mixture learning for video background subtraction," *IEEE-T-PAMI*, vol. 27, no. 5, pp. 827–832, May 2005.
[26] F. Luthon, A. Caplier, and M. Lievin, "Spatiotemporal mrf approach to video segmentation: application to motion detection and lips segmentation," *Signal Processing*, vol. 76, no. 1, pp. 61–80, July 1999.
[27] A. Caplier, L. Bonnaud, and J-M. Chassery, "Robust fast extraction of video objects combining frame differences and adaptative reference image," in *IEEE-C-ICIP*, September 2001, vol. 2, pp. 785–788.
[28] S. Geman and D. Geman, "Bayesian restoration of images," *IEEE-T-PAMI*, vol. 6, no. 6, pp. 721–741, 1984.
[29] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society*, vol. B-48, no. 3, pp. 259–302, 1986.
[30] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *CVIU*, vol. 80, no. 1, pp. 42–56, October 2000.
[31] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," *PIEEE*, vol. 83, no. 5, pp. 705–740, May 1995.
[32] T. Fromherz, P. Stucki, and M. Bichsel, "A survey of face recognition," *MML Technical Report, No 97.01, Dept. of Computer Science*, 1997.
[33] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE-T-PAMI*, vol. 24, no. 1, pp. 34–58, January 2002.
[34] E. Hjelmas and B. K. Low, "Face detection: a survey," *CVIU*, vol. 83, no. 3, pp. 236–274, September 2001.
[35] W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips, "Face recognition: A literature survey," in *UMD University of Maryland – TR4167R*, 2002.
[36] J. Yang, W. Lu, and A. Waibel, "Skin-color modeling and adaptation," *ACCV*, vol. 2, pp. 687–694, January 1998.
[37] J. C. Terrillon, M. N. Shirazi, H. Fukamachi, and S. Akamatsu, "Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images," in *IEEE-C-AFGR*, March 2000, pp. 54–61.
[38] R. Brunelli and T. Poggio, "Face recognition: features versus templates," *IEEE-T-PAMI*, vol. 15, no. 10, pp. 1042–1052, 1993.
[39] A. P. Pentland, B. Moghaddam, and T. E. Starner, "View-based and modular eigenspace for face recognition," in *IEEE-C-CVPR*, June 1994, pp. 84–91.
[40] V. Girondel, L. Bonnaud, and A. Caplier, "Hands detection and tracking for interactive multimedia applications," in *ICCVG*, September 2002, vol. 1, pp. 282–287.
[41] V. Girondel, "Détection de peau, suivi de tête et de mains pour des applications multimédia," *SIPT Master's Technical Report*, July 2002.
[42] D. Chai and K. N. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE-T-CSVT*, vol. 9, no. 4, pp. 551–564, June 1999.
[43] S. L. Dockstader and A. M. Tekalp, "On the tracking of articulated and occluded video object motion," *Real Time Imaging*, vol. 7, no. 5, pp. 415–432, October 2001.
[44] M. B. Capellades, D. Doermann, D. DeMenthon, and R. Chellappa, "An appearance based approach for human and object tracking," in *IEEE-C-ICIP*, September 2003, vol. 2, pp. 85–88.
[45] V. Girondel, A. Caplier, and L. Bonnaud, "Real time tracking of multiple persons by kalman filtering and face pursuit for multimedia applications," in *IEEE-S-SSIAI*, 2004, pp. 201–205.
[46] R. E. Kalman, "A new approach to linear filtering and prediction problems," *T-ASME*, vol. 82, pp. 35–45, March 1960.
[47] A. F. Bobick and A. D. Wilson, "A state based approach to the representation and recognition of gesture," *IEEE-T-PAMI*, vol. 19, no. 12, pp. 1325–1337, December 1997.
[48] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *IEEE-C-CVPR*, June 1992, pp. 379–385.
[49] Y. Guo, G. Xu, and S. Tsuji, "Understanding human motion patterns," in *IEEE-C-ICPR*, 1994, vol. B, pp. 325–329.
[50] V. Girondel, L. Bonnaud, A. Caplier, and M. Rombaut, "Static human body postures recognition in video sequences using the belief theory," in *IEEE-C-ICIP*, September 2005, vol. 2, pp. 45–48.
[51] V. Girondel, L. Bonnaud, and A. Caplier, "A belief theory based static posture recognition system for real-time video surveillance applications," in *IEEE-C-AVSS*, September 2005, pp. 10–15.
[52] Z. Hammal, A. Caplier, and M. Rombaut, "Classification d'expressions faciales par la théorie de l'évidence," in *LFA*, 2004, pp. 173–180.
[53] Z. Hammal, L. Couvreur, A. Caplier, and M. Rombaut, "Facial expression recognition based on the belief theory: comparison with different classifiers," in *13th ICIAP*, September 2005, pp. 743–752.
[54] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, pp. 191–234, 1994.
[55] P. Smets, "The transferable belief model for quantified belief representation," in *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol. 1*, D. M. Gabbay and P. Smets, Eds., pp. 267–301. Kluwer, Doordrecht, The Netherlands, 1998.
[56] A. Dempster, "A generalization of bayesian inference," *Journal of the Royal Statistical Society*, vol. 30, pp. 205–245, 1968.
[57] G. Shafer, "A mathematical theory of evidence," *Princeton University Press*, 1976.
[58] E. Salvador, A. Cavalarro, and T. Ebrahimi, "Shadow identification and classification using invariant color models," in *IEEE-C-ICASSP*, May 2001, pp. 1545–1548.
[59] P. C. Hernandez, J. Czyz, T. Umeda, F. Marques, X. Marichal, and B. Macq, "Silhouette based probabilistic 2d human motion estimation for real time applications," in *IEEE-C-ICIP*, September 2005.
[60] Z. Hammal, C. Massot, G. Bedoya, and A. Caplier, "Eyes segmentation applied to gaze direction and vigilance estimation," in *3rd ICAPR*, August 2005, pp. 236–246.

Vincent Girondel was born in Caen (France) in 1978. He graduated from the École Nationale Supérieure d'Électronique et de Radioélectricité de Grenoble (ENSERG) of the Institut National Polytechnique de Grenoble (INPG), France, in 2001. He obtained his Master's degree in Signal, Image, Speech Processing and Telecommunications from the INPG in 2002. He is currently a temporary teaching and research assistant at the ENSERG and at the Laboratoire des Images et des Signaux (LIS), and he is finishing his PhD at the LIS in Grenoble.

His research interests include human motion analysis from low-level to high-level interpretation, data fusion and video sequences analysis for real-time mixed reality applications (segmentation, tracking, interpretation…).

Laurent Bonnaud was born in 1970. He graduated from the École Centrale de Paris (ECP) in 1993. He obtained his PhD from the Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) and the Université de Rennes-1 in 1998. Since 1999 he is teaching at the Université Pierre Mendès-France (UPMF) in Grenoble and is a permanent researcher at the Laboratoire des Images et des Signaux (LIS) in Grenoble.

His research interests include segmentation and tracking, human motion and gestures analysis and interpretation.

Alice Caplier was born in 1968. She graduated from the École Nationale Supérieure des Ingénieurs Électriciens de Grenoble (ENSIEG) of the Institut National Polytechnique de Grenoble (INPG), France, in 1991. She obtained her Master's degree in Signal, Image, Speech Processing and Telecommunications in 1992 and her PhD from the INPG in 1995. Since 1997 she is teaching at the École Nationale Supérieure d'Électronique et de Radioélectricité de Grenoble (ENSERG) of the INPG and is a permanent researcher at the Laboratoire des Images et des Signaux (LIS) in Grenoble.

Her interest is on human motion analysis and interpretation. More precisely, she is working on the recognition of facial gestures (facial expressions and head motion) and the recognition of human postures.