A belief theory-based static posture recognition system for real-time videosurveillance applications

Vincent Girondel, Alice Caplier, Laurent Bonnaud

A Belief Theory-Based Static Posture Recognition System for Real-Time Video Surveillance Applications

V. Girondel A. Caplier L. Bonnaud

Laboratoire des Images et des Signaux (LIS)
Institut National Polytechnique de Grenoble (INPG), France
http://www.lis.inpg.fr

Abstract

This paper presents a system that can automatically recognize four different static human body postures for video surveillance applications. The considered postures are standing, sitting, squatting, and lying. The data come from the persons 2D segmentation and from their face localization. It consists in distance measurements relative to a reference posture (standing, arms stretched horizontally). The recognition is based on data fusion using the belief theory, because this theory allows the modelling of imprecision and uncertainty. The efficiency and the limits of the recognition system are highlighted thanks to the processing of several thousands of frames. A considered application is the monitoring of elder people in hospitals or at home. This system allows real-time processing.

1. Introduction

Human motion analysis is one of the most active research field in computer vision devoted to detecting, tracking and understanding people physical behavior. Human motion analysis has recently received a lot of attention [1, 2, 3]. This strong interest is driven by a wide spectrum of applications in various areas such as video surveillance, athletic performance analysis, video access control to sites, content-based video storage and retrieval, interactive mixed reality systems, perceptual human-computer interface [4] etc.

The video surveillance area covers applications where one or more people are being tracked over time and monitored for special actions. The strong need of smart video surveillance systems comes from the existence of security-sensitive areas such as banks, department stores, parking lots and borders. Surveillance cameras outputs in these places are often stored in video archives or recorded on tapes. Most of the time, these video data are only used "after the fact" mainly as an identification tool. The primary benefit fact that the camera is an active real-time processing media is therefore sometimes unused. The need is the real-time video analysis of sensitive places in order to alert police officers of a burglary in progress, or of the suspicious presence of a human wandering for a long time in a parking lot. As well as these obvious security applications, smart video surveillance can also be used to measure and control the traffic flow, compile consumer demographics in shopping malls, monitor elder people etc.

A study by Haritaoglu et al. [5] for the W^4 real-time visual surveillance system operates on monocular grey scale or on infrared video sequences. W^4 makes no use of color cues, instead it employs a combination of shape analysis and tracking to locate people and their body parts.

In the DARPA VSAM project, Collins et al. of the CMU designed a system for video-based surveillance [6]. Using multiple sensors, the system classifies and tracks multiple people and vehicles. With a star skeletonization procedure for people, they succeed in determining the gait and posture of a moving human being, classifying its motion between walking and running.

In [7], Nair and Clark describe an automated visual surveillance system that can classify human activities and detect suspicious events in a scene. This real-time system detects people in a corridor, tracks them and uses dynamic information to recognize their activities. Using a set of discrete and previously trained HMMs, it manages to classify people entering or exiting a room, and even mock break-in attempts. However, the system's false alarm rate is high since there are many other possible activities.

In this paper, we present a system that can automatically recognize in real-time four different static human body postures (standing, sitting, squatting and lying). It uses dynamic video sequence analysis information and data fusion using the belief theory. The TBM (Transferable Belief Model) was introduced by Smets in [8, 9]. It follows the works of Dempster [10] and Shafer [11]. The main advantage of the belief theory is the possibility to model imprecision and conflict. It is also not computationally expensive, compared with HMMs and, as doubt is taken into account, leads to a low false alarm rate. Here a possible application is

elder people monitoring, detecting for instance if someone has fallen down or have been sitting for too long.

The remainder of this paper is organized as follows. Section 2 presents an overview of the process with the pre-processing steps, the acquisition conditions and the data used in our method. Then, section 3 illustrates the main steps of the belief theory, the static posture recognition problem and the proposed solution. Section 4 shows the obtained classification results. Some details about the implemented system are given in section 5. Finally, section 6 concludes the paper and gives some perspectives.

2. Overview

The filmed environment consists in an indoor scene where people can enter one at a time. Our hypotheses are that each person is to stay approximately at the same distance of the static camera, is observed at least once in the reference posture ("Da Vinci Vitruvian Man posture", see Fig. 1) and is not completely occluded. Before the posture recognition step, there are three pre-processing steps. The first step is the **segmentation** of the people. It is performed by an adaptive background removal algorithm [12]. Then the rectangular bounding box (*RBB*), the principal axes box (*PAB*) which is a box whose directions are given by the principal axes of the person shape, and the gravity center are computed (see Fig. 1). The second step is the temporal **tracking** of people. The third step is the **face and hands localization** of each person [13].

Three distances are computed (see Fig. 1): D_1 the vertical distance from the face center to the person *RBB* bottom, D_2 the distance from the face center to the person *PAB* center (gravity center) and D_3 the person *PAB* semi great axe length. Each distance D_i is normalized with respect to the corresponding distance D_i^{ref} obtained when the person is in the reference posture in order to take into account the inter-individual variations of heights. The measurements are noted $r_i = D_i/D_i^{ref}$ $(i = 1 \dots 3)$. Fig. 2 illustrates the variations of r_1 for people in the same postures succession: reference posture, sitting, standing, squatting, standing, lying, standing, sitting, standing and lying. Although the time taken to sit down, to stand up or just to be in a static posture is not the same, the different patterns and levels for the four static postures are clearly visible. The static postures sequence for the third person (Vincent) is shown at the bottom of Fig. 2 as the corresponding hypothesis label H_i $(i = 1 \dots 4)$. H_1, H_2, H_3 and H_4 correspond to static postures and H_0 refers to postures occurring during the transitions steps which are unknown postures.

The aim of the work presented here is to design a recognition system of static human body postures based on the fusion of the r_i measurements using the belief theory.

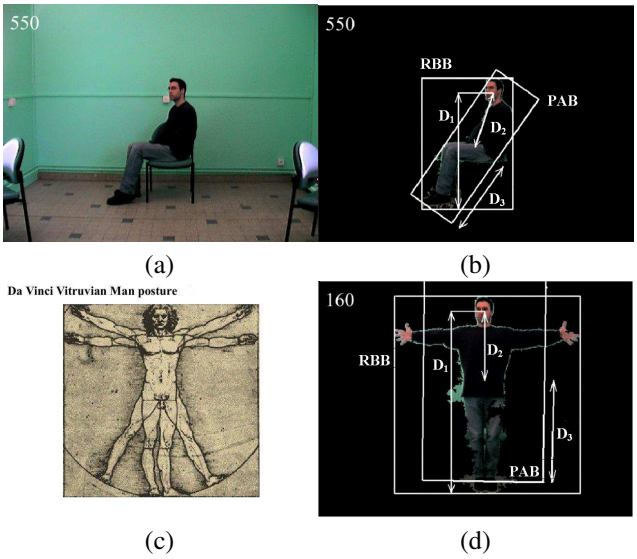

Figure 1: Examples of distances for two postures. (a, b) sitting posture, (c, d) reference posture.

3. Belief theory

The belief theory approach needs the definition of a universe Ω composed of N disjunctive hypotheses H_i. If the hypotheses are exhaustive, Ω is a closed universe. In this paper, we consider an open universe, as all possible human body postures can not be classified in the following four static postures: standing (H_1), sitting (H_2), squatting (H_3), and lying (H_4). We add an hypothesis for the unknown posture class (H_0). 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 favoring 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:

$$\begin{array}{rcl} m : 2^\Omega & \longrightarrow & [0;1] \\ A & \longmapsto & m(A) \end{array}$$

with $\sum_{A \in 2^\Omega} m(A) = 1$.

3.1. Modelling

A model has to be defined for each r_i in order to associate a belief mass to each subset A, depending on the value of r_i. Two different models types are used (see Fig. 3). The first model type is used for r_1 and the second for r_2 and r_3.

Considering r_1 measurement, the first model type is based on the idea that the lower the face of a person is located, the closer the person is from the lying posture. On the

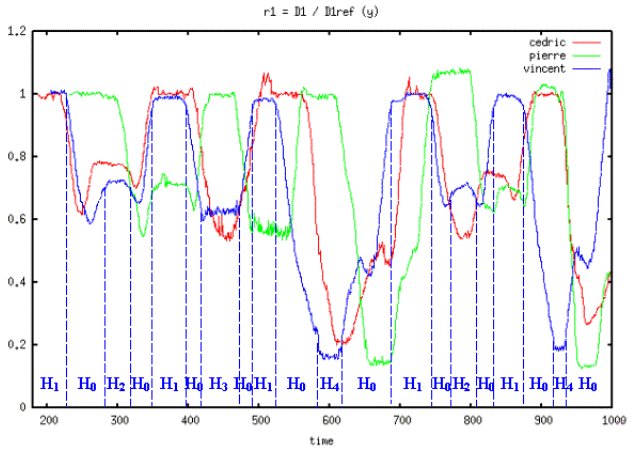

Figure 2: r_1 variations for 3 different persons.

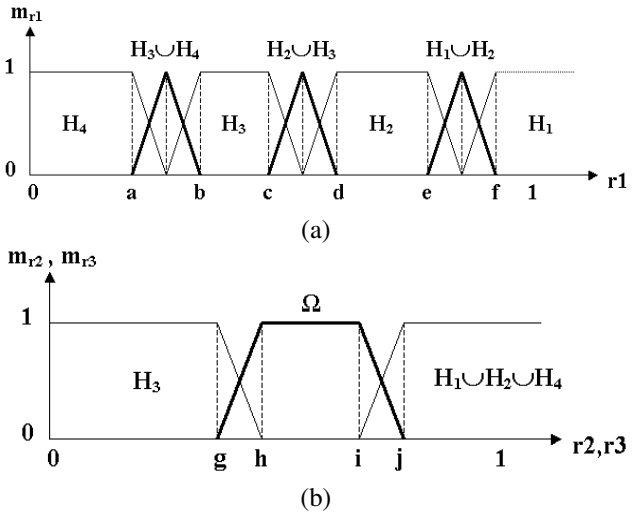

Figure 3: Belief models (a) m_{r_1}, (b) m_{r_2}, m_{r_3}. H_i defines recognized posture(s).

opposite, the higher the face is located, the closer the person is from the standing posture. Depending on the value of r_1, either a single posture is recognized and its belief mass is set to 1, or non-null belief masses are given to a single posture and to the union of two postures. The second case illustrates the modelling of the imprecision and uncertainty in the models.

Considering r_2 and r_3 measurements, 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, h, i and 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. Thanks to data fusion, the recognition of static human body postures is then possible.

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

$$m_{r_{ij}} = m_{r_i} \oplus m_{r_j} \quad (1)$$

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

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. It usually happens when some of the r_i are in the transition zones of the models.

Many tests are performed to find the fourteen most suitable thresholds (a-f for r_1, g-j for r_2 and g-j for r_3), for the belief models, (see Fig. 3). The most suitable thresholds yield a minimum of conflict. The statistics of the three measurements r_i are computed (minima, maxima, means and standard deviations) over a training set of ∼5000 frames (see section 4). In fact, one of hardest step in the belief theory is to find a modelling way that leads to a minimum of conflicts.

The orthogonal sum is associative and commutative, so the order to fusion the belief mass distributions does not matter. Here, the chosen order to compute $m_{r_{123}}$ is :

$$m_{r_{23}} = m_{r_2} \oplus m_{r_3}$$
$$m_{r_{123}} = m_{r_1} \oplus m_{r_{23}}$$

3.3. 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. The choice is based on the resulting belief mass distribution. A criterion $Crit$ defined on the final belief mass distribution is generally optimized to choose the classification result $\hat{A}$:

$$\hat{A} = \arg\max_{A \in 2^\Omega} Crit(A).$$

Note that $\hat{A}$ may not be a singleton but a union of several hypotheses or even the empty set. There are usual criteria used to make a decision: the belief mass, the belief noted

Bel, the plausibility noted Pl etc. With the paper notations, these three criteria are defined by the following expressions:

$$Crit(A) = m_{r_{123}}(A) \quad (3)$$

$$Crit(A) = Bel(A) = \sum_{B \in 2^\Omega, B \subset A} m_{r_{123}}(B) \quad (4)$$

$$Crit(A) = Pl(A) = \sum_{B \in 2^\Omega, A \cap B \neq \emptyset} m_{r_{123}}(B) \quad (5)$$

In this paper, the hypothesis H_0 is chosen if the classification result is the empty set $\emptyset$, i.e. $Crit(\emptyset)$ is maximum.

By definition, the Bel and Pl criteria (resp. (4) and (5)) yield greater values for subsets of 2^Ω with numerous elements. Therefore, it can be useful to compute these criteria only for singletons and for the empty set, in order to limit the classification result to a single hypothesis.

Two choices are possible, the first is to make the decision on all subsets of 2^Ω, the other, only on singletons and on the empty set. Section 4 gives the results for both cases.

3.4. Example

For instance, with the following distributions:

$$m_{r_1}(H_2 \cup H_3) = 0.8 \qquad m_{r_{23}}(H_3) = 0.9$$
$$m_{r_1}(H_2) = 0.2 \qquad m_{r_{23}}(\Omega) = 0.1$$

According to the following conjunction table,

$m_{r_1} \backslash m_{r_{23}}$	H_3	Ω
$H_2 \cup H_3$	H_3	$H_2 \cup H_3$
H_2	$\emptyset$	H_2

the belief mass of each resulting subset is:

$$m_{r_{123}}(H_3) = 0.72 \qquad m_{r_{123}}(H_2 \cup H_3) = 0.08$$
$$m_{r_{123}}(\emptyset) = 0.18 \qquad m_{r_{123}}(H_2) = 0.02$$

If we compute the different criteria for the previously obtained subsets whose belief masses are different from 0:

- Eq. (3) yields H_3
- Eq. (4) and (5) yield $H_2 \cup H_3$

Because we have $Bel(H_2 \cup H_3) = Pl(H_2 \cup H_3) = 0.82$, $Bel(H_2) = 0.08$, $Bel(H_3) = 0.72$, $Pl(H_2) = 0.1$ and $Pl(H_3) = 0.8$.

If the different criteria are computed only for singletons and for the empty set:

- Eq. (3), (4) and (5) yield H_3

3.5. Implementation

The major problem of the belief theory is the combinatory explosion when computing the orthogonal sum(s). This problem can be alleviated by a clever implementation. The solution is to code each hypothesis by a power of two. Here, the choice is: $H_0 = 0$, $H_1 = 1$, $H_2 = 2$, $H_3 = 4$ and $H_4 = 8$. The conjunction code for two combinations of hypotheses is the $logical\ and$ of their binary coding: $(H_1 \cup H_2) \cap H_1 = 11 \cap 01 = 01 = H_1$. One can clearly see that the belief mass of a conflict will be associated to H_0: $H_1 \cap H_2 = 01 \cap 10 = 00 = H_0$. The orthogonal sum can be computed for all subsets at the same time. In our case, to compute the orthogonal sum $m_{r_{ij}}$ of the belief mass distributions m_{r_i} and m_{r_j}, we compute:

```
For i from 0 to 15 (15=Card(2^Ω)-1)
 If m_{r_i}[i]≠0 (to avoid unnecessary For loops)
 For j from 0 to 15
  m_{r_ij}[i & j]+=m_{r_i}[i].m_{r_j}[j]
```

m_{r_i}`[i]` defines the m_{r_i} belief mass distribution value of the subset of 2^Ω coded by the binary code `i`. At the beginning, each belief mass distribution is initiated with 0. Then according to the different r_i values and the belief models, the belief masses of each subset are defined. Then, the $logical\ and$ `i & j` computes the conjunction product belief mass value between the respective subsets `i` of m_{r_i} and `j` of m_{r_j}. Loops begin from 0 because m_{r_i} and/or m_{r_j} belief mass distributions can have a non-null elementary belief mass on the empty set, i.e. a conflict, if they result from another orthogonal sum.

4. Results

In order to see the static posture recognition results, two sets of video sequences are used, a training set and a test set. The computation of the statistics of the normalized distance measurements is performed on the training set video sequences and the test step allows to see the robustness and the performances of the recognition system on non previously used video sequences.

The training set consists in twelve different video sequences representing ∼13000 frames. Six different people have been filmed twice in the same ten successive postures. People were of various heights, between 5.2 ft and 6.4 ft, in order to take into account the variability of heights and improve the robustness of the algorithm. The constraints were to be in "natural" postures in front of the camera.

The test set consists in twelve other video sequences representing ∼20000 frames. Six other people have been filmed, also twice, in different successive postures. People were not the same as those in the training step and were also of various heights. In order to test the limits of the system,

people were allowed to move the arms, sit sideways and even be in postures that do not often occur in everyday life.

We present the classification results obtained for two different classifiers. The first classifier named C_1 uses (3). The criterion computation is done for every subset of 2^Ω, i.e. not only singletons. The classification result for C_1 is therefore the subset of 2^Ω with the **maximum belief mass**.

As we want to recognize a single posture, it can be useful to compute a criterion only for singletons and for the empty set. The classifier is then forced to choose either a singleton or the unknown posture. The classification result is:

$$\hat{A} = \arg \max_{A \in 2^\Omega, Card(A)<2} Crit(A)$$

with the following definitions: $Bel(\emptyset) = m_{r_{123}}(\emptyset)$ and $Pl(\emptyset) = m_{r_{123}}(\emptyset)$. The non-considered subsets belief masses are not taken into account for the decision if we use (3) or (4), whereas they are taken into account if (5) is used, by definition of the Pl. The second classifier named C_2 shows the obtained classification results using (5) computed only for singletons and for the empty set. Hence the classification result for C_2 is, between singletons and the empty set, the subset with the **maximum plausibility**. The results are very similar to those obtained with the classifiers using (3) or (4), but they are a little better.

Results for C_1 and C_2 are computed on temporal parts of the video sequences where the global body posture is static, at least for the person's trunk. They represent the processing of $\sim$16000 frames out of 33000.

4.1. Training step

Training step recognition rates: Training recognition rates for the two classifiers C_1 and C_2 are available in Tables 1 and 2. Columns show the real posture and lines the postures recognized by the system.

Table 1: C_1 training confusion matrix

Syst\H	H_1	H_2	H_3	H_4
H_0	0%	0.1%	0%	0%
H_1	**100%**	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_4	0%	0%	0%	**100%**

As the belief models thresholds results from the r_i statistical characteristics computation over this set of video sequences, results are very good. There is only 0.1% of occurring conflicts on more than $\sim$5000 frames. There are no problems to recognize the standing or the lying postures. The sitting and the squatting postures are also well recognized even if there is a little doubt between both.

Table 2: C_2 training confusion matrix

Syst\H	H_1	H_2	H_3	H_4
H_0	0%	0.1%	0%	0%
H_1	**100%**	0%	0%	0%
H_2	0%	**97.2%**	1.5%	0%
H_3	0%	2.7%	**98.5%**	0%
H_4	0%	0%	0%	**100%**

In the case of the C_2 classifier, results are even better. There are always no problems for the standing or the lying postures. By computing the plausibilities only for singletons, the classifier is forced to choose between H_2 and H_3 instead of choosing $H_2 \cup H_3$. That leads to a better recognition in more than half of the cases.

The average recognition rates for the two classifiers are the following ones: C_1: **97.7%**, C_2: **98.9%**.

4.2. Test step

Test step recognition rates: Test recognition rates for the two classifiers C_1 and C_2 are available in Tables 3 and 4. Columns show the real posture and lines the postures recognized by the system.

Table 3: C_1 test confusion matrix

Syst\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_4	0%	0%	0%	**100%**

For the C_1 classifier, there are more recognition errors but the results show a good global recognition rate. There are always no problems to recognize 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 reason is everybody does not sit and/or squat the same way, hands on knees or touching ground, back bent or straight etc. That fact yields more conflicts, whose number is 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 4: C_2 test confusion matrix

Syst\H	H_1	H_2	H_3	H_4
H_0	0%	10.2%	5.0%	0%
H_1	**99.9%**	0.4%	0%	0%
H_2	0.1%	**71.6%**	30.9%	0%
H_3	0%	17.8%	**64.1%**	0%
H_4	0%	0%	0%	**100%**

For the C_2 classifier, on the one hand, results are again improved because the recognition rates H_i vs H_i are better, on the other hand there are also more recognition errors between H_2 and H_3.

The average recognition rates for the two classifiers are the following ones: C_1: **78.1%**, C_2: **83.9%**.

5. Implemented system

Video sequences are acquired with a Sony $DFW-VL500$ camera, in the YC_bC_r 4:2:0 format at 30 fps and in 640×480 resolution. The results are obtained at a frame rate of $\sim$14 fps on a PC running at 3.2 GHz. Code has been implemented in unoptimized C++. A lot of video sequences have been tested for each step of the process, representing several thousands of images. Real-time processing can be easily achieved by optimizing the C++ code and by reducing the resolution to 320×240.

6. Conclusion and perspectives

We presented in this paper a method based on the belief theory to recognize four static human body postures with a few number of normalized distance measurements. This method has shown good recognition results and is fast enough to allow real-time processing.

This system could be used to monitor elder 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 elder people, their postures are likely to resemble the training set ones. In these conditions, the system should be reliable enough to succeed in this monitoring as the training recognition rates are very good.

The major problem of this method is the fact that a person has to do the reference posture again if the distance to the camera changes significantly. If this system is used for indoor video surveillance inside a room, there should be no problems if the initialization is correctly done. Otherwise, one solution under study is to use a stereo camera that can measure the depth and use this information to normalize the distances computed on the person's mask. Another problem is the posture recognition during the transition between two static postures. We plan to enhance the method by adding a dynamic analysis of the measurements temporal evolution. This should greatly improve the recognition results. To justify this positive statement, an interesting point can be seen on Figure 2. When a person is sitting down, the variation of r_1 has a characteristic pattern: it decreases before increasing again because the person bends forward instead of sitting straight downward (this also happens when a person stands up). That is a point for a dynamic analysis which could lead to real-time recognition of dynamic postures and actions recognition like standing up, sitting down etc.

References

[1] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[2] J. J. Wang and S. Singh, "Video analysis of human dynamics - a survey," *Real-Time Imaging*, vol. 9, pp. 321–346, 2003.

[3] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern Recognition*, vol. 36, no. 3, pp. 585–601, 2003.

[4] Website of the SIMILAR European Network of Excellence, "http://www.similar.cc/," .

[5] I. Haritaoglu, D. Harwood, and L. Davis, "Who, when, where, what: A real time system for detecting and tracking people," in *IEEE International Conference on Automatic Face and Gesture Recognition*, April 1998, pp. 222–227.

[6] 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," in *CMU-RI-TR*, 2000.

[7] V. Nair and J. J. Clark, "Automated visual surveillance using hidden markov models," in *VI02*, 2002, p. 88.

[8] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, pp. 191–234, 1994.

[9] 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.

[10] A. Dempster, "A generalization of bayesian inference," *Journal of the Royal Statistical Society*, vol. 30, pp. 205–245, 1968.

[11] G. Shafer, "A mathematical theory of evidence," *Princeton University Press*, 1976.

[12] A. Caplier, L. Bonnaud, and J-M. Chassery, "Robust fast extraction of video objects combining frame differences and adaptative reference image," in *IEEE International Conference on Image Processing*, September 2001.

[13] V. Girondel, L. Bonnaud, and A. Caplier, "Hands detection and tracking for interactive multimedia applications," in *International Conference on Computer Vision and Graphics*, September 2002, pp. 282–287.