Belief theory-based classifiers comparison for static human body postures recognition in video

V. Girondel, L. Bonnaud, A. Caplier, and M. Rombaut

***Abstract*—** This paper presents various classifiers results from a system that can automatically recognize four different static human body postures in video sequences. The considered postures are standing, sitting, squatting, and lying. The three classifiers considered are a naïve one and two based on the belief theory. The belief theory-based classifiers use either a classic or restricted plausibility criterion to make a decision after data fusion. The data come from the people 2D segmentation and from their face localization. Measurements consist in distances relative to a reference posture. The efficiency and the limits of the different classifiers on the recognition system are highlighted thanks to the analysis of a great number of results. This system allows real-time processing.

***Keywords*—**Belief theory, classifiers comparison, data fusion, human motion analysis, real-time processing, static posture recognition.

I. Introduction

HUMAN motion analysis is an important area of research in computer vision devoted to detecting, tracking and understanding people physical behavior. This strong interest is driven by a wide spectrum of applications in various areas such as smart video surveillance, interactive virtual reality systems, athletic performance analysis, perceptual human-computer interface (HCI) etc.

The next generation of HCIs will be multimodal, integrating the analysis and the recognition of human body postures and actions as well as speech and facial expressions analysis [1]. Behavior understanding is the ability to analyze human action patterns, and to produce high-level interpretation of these patterns. For many applications, it is necessary to be able to recognize particular human body postures. For example, it is important to know, in a video surveillance system of old people, if a person has fallen down and is lying motionless. The action recognition problem has recently received a lot of attention [2], [3], [4].

Human action recognition can be divided into dynamic and static recognition. In a lot of methods, a comparison between recorded information and the current image is done. Information may be templates [5], transformed templates [6], normalized silhouettes [7], or postures [8]. The aim of static recognition is mainly to recognize various postures, e.g., pointing, standing and sitting, or specially defined postures. Sul *et al.* [5] design an interactive Karaoke system where the postures of the subject are used to trigger and control the system. Templates are also used in the work of Oren *et al.* [6]. In an off-line process, they segment pedestrians and generate a common template based on Haar wavelets. In an on-line process, the template is compared to various parts of the image to find pedestrians.

In this paper, we present a method using classifiers to recognize four static human body postures. Static recognition is based on information obtained by dynamic sequence analysis. For instance we try to recognize the standing posture but not the standing up motion. We compare the obtained recognition rates using either a naïve classifier or classifiers based on data fusion using the belief theory. The belief theory was introduced by Shafer [9], after the first developments made by Dempster [10]. This model is an extension of the probabilities theory. The TBM (Transferable Belief Model) was really introduced in [11], [12]. The advantage of this theory is the possibility to model imprecision and conflict. To our knowledge, belief theory has not yet been used for human posture recognition. Therefore we describe a supervised classification system of static human body postures using the belief theory.

Manuscript received February 15, 2005.

II. 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 and is observed at least once in the reference posture ("Da Vinci Vitruvian Man posture", see Fig. 1 (c)). Before the posture recognition step, there are three preprocessing steps. The first step is the **segmentation** of people. It is performed by an adaptive background removal algorithm [13]. Then the Vertical Bounding Box (VBB), the Principal Axes Box (PAB) which is a box whose directions are given by the principal axes of the person shape, see Fig. 1 (b, d), and the gravity center are computed . The second step is the temporal **tracking** of people. The third step is the **face and hands localization** of each person. It uses skin detection based on color information with an adaptive thresholding [14]. Four distances are then computed, see Fig. 1 (b, d): D_1 is the vertical distance from the face center to the VBB bottom, D_2 is the VBB height, D_3 is the distance from the face center to the PAB center (gravity center) and D_4 is the PAB semi great axe length. The work here aims at designing a recognition system based on these distances.

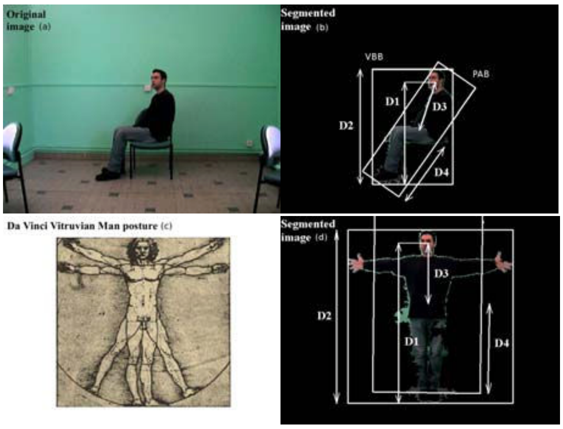

Fig. 1. Examples of distances for two postures: the sitting posture (a, b) and the reference posture (c, d).

III. Belief theory

The belief theory approach needs the definition of a universe Ω composed of N disjunctive hypotheses H_i. In this paper, the hypotheses are the four static postures, plus one for the unknown posture class: unknown (H_0), standing (H_1), sitting (H_2), squatting (H_3), and lying (H_4). 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:

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

with $\sum_{A \in 2^\Omega} m\,(A) = 1$. The measurements are the four distances D_i $(i = 1 \ldots 4)$ presented in Section II. Each distance is normalized with respect to the corresponding distance obtained when the person is in the reference posture in order to take into account the inter-individual variations of heights: $r_i = D_i / D_i^{ref}$ $(i = 1 \ldots 4)$. 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. The postures sequence for the third person (Vincent) is shown at the bottom of Fig. 2 as the corresponding hypothesis label $H_{1 \ldots 4}$. H_1, H_2, H_3, and H_4 correspond to static postures and H_0 refers to the postures occurring during the transition steps or unknown postures.

A. Modeling

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. r_1 and r_3 are based on the face position whereas r_2 and r_4 are not. Multiple independent measurements are used to improve the system reliability. Most of the time, i.e. when a person's arms are not raised, r_1 and r_2 (respectively r_3 and r_4) vary in the same range. Therefore two different models are used (see Fig. 3). The first model is used for r_1 and r_2 and the second model for r_3 and r_4. The first model 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 opposite, the higher the face is located, the closer the person is from the standing posture. The second model 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. In the case of raised arms, measurements using face localization are more robust. All Fig. 3 thresholds (a-h) are obtained after a human expertise over a training set of twelve different video sequences (see Section IV). The thresholds are different for each r_i.

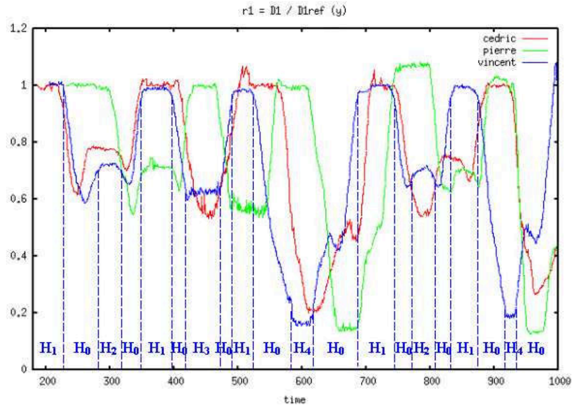

Fig. 2. r_1 variations for three different persons.

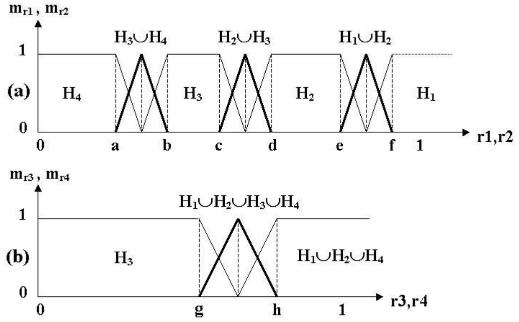

Fig. 3. Belief models m_{r_1}, m_{r_2} (a), m_{r_3}, m_{r_4} (b).
H_i defines recognized posture(s).

B. Data fusion

The aim is to obtain a belief mass distribution $m_{r_{1234}}$ that takes into account all available information (the belief mass distributions of every 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 A subset of 2^Ω:

$$m_{r_{ij}} = m_{r_i} \oplus m_{r_j} \tag{1}$$

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

The orthogonal sum of m_{r_1} and m_{r_2} on one side, m_{r_3} and m_{r_4} on the other side, yields $m_{r_{12}}$ and $m_{r_{34}}$. $m_{r_{1234}}$ is obtained through their orthogonal sum. For instance, take the two following belief mass distributions $m_{r_{12}}$ and $m_{r_{34}}$:

$$\begin{aligned} m_{r_{12}}(H_2 \cup H_3) = 0.8 \quad & m_{r_{34}}(H_3) = 0.9 \\ m_{r_{12}}(H_2) = 0.2 \quad & m_{r_{34}}(H_1 \cup H_2 \cup H_3 \cup H_4) = 0.1 \end{aligned}$$

According to the following table where $\emptyset$ is the empty set:

$r_{12} \backslash r_{34}$	H_3	$H_1 \cup H_2 \cup H_3 \cup H_4$
$H_2 \cup H_3$	H_3	$H_2 \cup H_3$
H_2	$\emptyset$	H_2

The belief mass of each resulting subset is:

$$\begin{aligned} m_{r_{1234}}(H_3) = 0.72 \quad & m_{r_{1234}}(H_2 \cup H_3) = 0.08 \\ m_{r_{1234}}(\emptyset) = 0.18 \quad & m_{r_{1234}}(H_2) = 0.02 \end{aligned}$$

In case when $m_{r_{1234}}(\emptyset) \neq 0$, there is a **conflict**, which means that modeling rules give contradictory results. It usually happens when some of the r_i correspond to wrong posture(s) or are in the transition zones of the models. For instance, grouping H_2 with H_1 and H_4 in the model (b) of Fig. 3 prevents conflicts when a person sits with a (or both) raised hand(s). In fact, the most difficult part of the belief theory is the measurements modeling so that data fusion yields a minimum of conflicts.

C. Decision

The decision is the final step of the process. Once all the belief mass distributions have been combined into a single one *m*, 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 *m* is 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 ($\emptyset$) if the criterion is maximum for it. There are usual criteria used to make a decision:

$$\text{belief mass: } Crit(A) = m(A) \quad (3)$$

$$\text{belief: } Crit(A) = Bel(A) = \sum_{B \in 2^\Omega, B \subset A} m(B) \quad (4)$$

$$\text{plausibility: } Crit(A) = Pl(A) = \sum_{B \in 2^\Omega, A \cap B \neq \emptyset} m(B) \quad (5)$$

For the example of subsection III.*B* the decision is H_3 if we use (3). If we use (4) or (5), the decision is $H_2 \cup H_3$ because $Bel(H_2 \cup H_3) = Pl(H_2 \cup H_3) = 0.82$. We have also $Bel(H_2) = 0.08$, $Bel(H_3) = 0.72$, $Pl(H_2) = 0.1$, and $Pl(H_3) = 0.8$. Here, if $Crit(\emptyset)$ is the highest, the hypothesis H_0 is chosen.

D. Classifiers comparison

Three classifiers are tested, all based on the belief models defined in subsection III.*A*. The first one is a naïve classifier, named C_1: each r_i value corresponds to a given hypothesis or combination of hypotheses. Fig. 4 illustrates the naïve models S_{r_i} which derive from the belief models m_{r_i} of Fig. 3. For each r_i, we increment the corresponding posture(s) score(s). At the end, the posture with the highest score is the recognized posture. If several postures have the same score, the chosen posture is the first one appearing in Ω, since postures are ordered according to *a priori* likelihood.

Belief theory-based classifiers using (3), (4) or (5) are tested and yield very similar results ((5) being slightly better). In fact, most of the time, the recognized postures are either a single posture (singleton) or an unknown posture (conflict). Recognition of a combination of hypotheses occurs very rarely. Therefore we focus on singletons because we are looking for single postures. We choose H_0 if $Crit(\emptyset)$ is the highest or if $Crit(A)$ is the highest and $Card(A) > 1$. For singletons, (3) and (4) are equivalent. The plausibility classifier, named C_2, is the second classifier and uses (5). Plausibilities are computed only for singletons of Ω and for H_0. By extension, we define $Pl(H_0)$ as

$$Pl(H_0) = m_{r_{1234}}(\emptyset) + \sum_{A \in 2^\Omega, Card(A) > 1} m_{r_{1234}}(A).$$

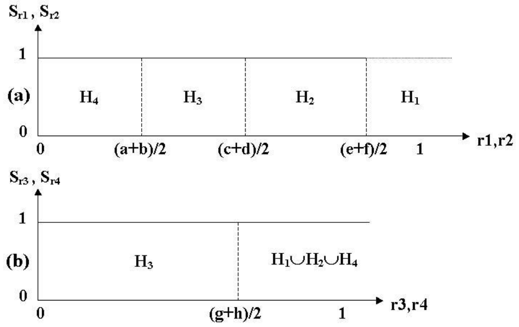

Fig. 4. Naïve models S_{r_1}, S_{r_2} (a), S_{r_3}, S_{r_4} (b).
H_i defines recognized posture(s).

C_2 shows better results than the classifiers using (3) or (4).

The third classifier, C_3, is a restricted plausibility-based classifier with $Pl(H_0) = 0$. C_3 is useful to compare naïve data fusion efficiency with belief theory-based data fusion efficiency and does not take into account H_0.

E. Implementation

The major problem of the belief theory is the combinatory explosion. 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 (2) can be computed for all subsets A at the same time.

IV. Results

A. Implemented system and computing time

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 approximately 11 fps on a low-end PC running at 1.8 GHz. Real-time processing could be easily achieved by optimizing the C++ code and by reducing the resolution to 320 × 240.

B. Training and test steps

For the training step, six different persons have been filmed twice in the same ten successive postures. The constraints were to be in "natural" postures in front of the camera. For the test step, six other people have been filmed twice, in different successive postures. People were free this time, no constraints have been given in the way to do each posture (move the arms, sit sideways etc).

Results for the three classifiers C_1, C_2 and C_3 are computed on temporal parts of the sequences where the global body posture is static, at least for the person's trunk. They represent the processing of approximately 15000 frames over 33000. The recognition rates are available in Tables I, II and III. Columns show the real posture and lines the postures recognized by the system. The rates on the left correspond to the training step and those on the right to the test step.

Training step recognition rates: As the models thresholds have been determined by expertise, results are excellent for all tested classifiers, except for squatting in C_2. The first reason is everybody does not squat the same way, hands on knees or touching ground, back bent or straight etc. That fact yields a lot of conflicts. Then, the human expertise determined thresholds near the squatting posture are surely not optimal. Note that C_2 generally recognizes either the right posture or the unknown posture whereas the other classifiers can only make errors (columns for H_2 and H_3). The average recognition rates are **95.7%** for C_1, **86.5%** for C_2, and **97.4%** for C_3.

TABLE I

C_1 NAÏVE CLASSIFIER CONFUSION MATRIX (TRAINING%/ TEST%)

Syst$\backslash H$	H_1	H_2	H_3	H_4
H_1	**100/ 100**	1.4/ 7.4	0/ 17.4	0/ 0
H_2	0/ 0	**97.6/ 85.8**	14.9/ 27	0/ 0
H_3	0/ 0	1.0/ 6.7	**85.1/ 55.6**	0/ 0
H_4	0/ 0	0/ 0	0/ 0	**100/ 100**

TABLE II

C_2 PLAUSIBILITY CLASSIFIER CONFUSION MATRIX (TRAINING%/ TEST%)

Syst$\backslash H$	H_1	H_2	H_3	H_4
H_0	0/ 0	2.4/ 22.5	48.5/ 36.3	0.5/ 0
H_1	**100/ 100**	0/ 0	0/ 0	0/ 0
H_2	0/ 0	**97.3/ 75.5**	2.3/ 24.9	0/ 0
H_3	0/ 0	0.3/ 2.0	**49.2/ 38.8**	0/ 0
H_4	0/ 0	0/ 0	0/ 0	**99.5/ 100**

TABLE III

C_3 RESTRICTED PLAUSIBILITY-BASED CLASSIFIER CONFUSION MATRIX (TRAINING%/ TEST%)

Syst$\backslash H$	H_1	H_2	H_3	H_4
H_1	**100/ 100**	0/ 0.9	0/ 5.6	0/ 0
H_2	0/ 0	**98.8/ 92.1**	9.2/ 33.9	0/ 0
H_3	0/ 0	1.2/ 7.0	**90.8/ 60.5**	0/ 0
H_4	0/ 0	0/ 0	0/ 0	**100/ 100**

Test step recognition rates: There are more recognition errors but the results show a good global recognition rate. For all classifiers, there are no problems to recognize the standing or the lying posture. The sitting posture is quite well recognized whereas squatting is confused with sitting (or unknown for C_3) when people have their arm(s) raised over their head. Results of C_1 and C_3 are similar, but C_3 performs better, showing more efficient data fusion thanks to the belief theory. We can see that squatting is more often confused with sitting in C_3 than in C_1, making errors on a ”closer” posture. In C_2, most of unknown postures are conflicts so the system detects a recognition problem instead of making a wrong choice. The average recognition rates are **85.3%** for C_1, **78.6%** for C_2, and **88.1%** for C_3.

V. CONCLUSION AND PERSPECTIVES

We presented a method based on the belief theory to recognize four static human body postures with a few number of normalized distances and compared the recognition results of three classifiers. The recognition rates show the efficiency of data fusion using the belief theory compared to the naïve data fusion although the models thresholds, determined by human expertise, are not robust enough (only for the squatting posture) to allow a good squatting posture recognition when people are allowed to move their arms. This method has shown good recognition results and is fast enough to allow real-time processing.

The major problem of this method is the fact that a person must do the reference posture again if the distance to the camera changes significantly. One solution could be the use of a stereo camera that can measure the depth and use this information to normalize the distances computed on the person 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 temporal evolution of the distances. This should greatly improve the recognition results. To justify this statement, an interesting point can be seen on Fig. 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 recognition of dynamic postures and actions recognition like standing up, sitting down etc. Further work will therefore deal with dynamic human body posture recognition.

REFERENCES

[1] URL of the website of the SIMILAR European Network of Excellence, “http://www.similar.cc/,” .

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

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

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

[5] C. W. Sul, K. C. Lee, and K. Wohn, “Virtual stage: a location-based karaoke system,” *IEEE Multimedia*, vol. 5, no. 2, pp. 42–52, 1998.

[6] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, “Pedestrian detection using wavelet templates,” in *Computer Vision and Pattern Recognition*, 1997, pp. 193–199.

[7] I. Haritaoglu, D. Harwood, and L. Davis, “Ghost: A human body part labeling system using silhouettes,” in *International Conference on Computer Vision and Pattern Recognition*, 1998, pp. 77–82.

[8] L. Campbell and A. Bobick, “Using phase space constraints to represent human body motion,” *International Workshop on Automatic Face and Gesture Recognition*, 1995.

[9] G. Shafer, “A mathematical theory of evidence,” *Princeton University Press*, 1976.

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

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

[12] 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.

[13] 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.

[14] 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.