Static human body postures recognition in video sequences using the belief theory

Vincent Girondel, Laurent Bonnaud, Alice Caplier, Michèle Rombaut

STATIC HUMAN BODY POSTURES RECOGNITION IN VIDEO SEQUENCES USING THE BELIEF THEORY

Vincent Girondel, Laurent Bonnaud, Alice Caplier, Michèle Rombaut

Laboratoire des Images et des Signaux (LIS),
961, rue de la Houille Blanche, BP 46, 38402 Saint Martin d'Hères Cedex, France
vgironde, bonnaud, caplier, rombaut@lis.inpg.fr

ABSTRACT

This paper presents 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 recognition is based on data fusion using the belief theory. The data come from the persons 2D segmentation and from their face localization. It consists in distance measurements relative to a reference posture ("Da Vinci posture": standing, arms stretched horizontally). The segmentation is based on an adaptive background removal algorithm. The face localization process uses skin detection based on color information with an adaptive thresholding. The efficiency and the limits of the recognition system are highlighted thanks to the analysis of a great number of results. This system allows real-time processing.

1. 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 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 analyse 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, for a video surveillance system of old people, it is important to know if the 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] designed 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 the belief theory 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. 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 by Smets in [11]. The advantage of this theory is the possibility to model imprecision and conflict. It has shown good results compared to other classifiers. 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 based on data fusion using the belief theory.

2. OVERVIEW

The filmed environment consists in an indoor scene where persons can enter one at a time. Our hypothesis are that each person is to stay approximately at the same distance of the static camera and be observed at least once in a reference posture. Before the posture recognition step, there are three pre-processing steps. The first step is the **segmentation** of the persons. It is performed by an adaptive background removal algorithm [12]. 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, and the gravity center are computed. The second

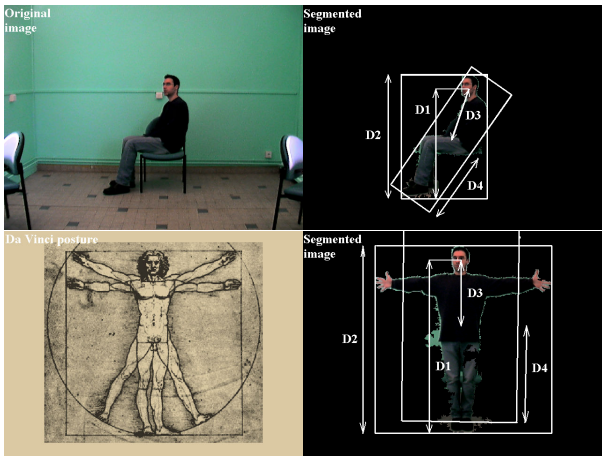

Fig. 1. Parameters for two postures: the sitting posture (top) and the reference posture (bottom).

step is the temporal **tracking** of persons. The third step is the **face and hands localization** of each person [13]. Four distances are then computed: D_1 is the vertical distance from the face center ordinate to the bottom's ordinate of the *VBB*, D_2 is the *VBB*'s height, D_3 is the distance from the face center to the *PAB*'s center (gravity center) and D_4 is the *PAB*'s semi great axe length. The aim of the work presented here is to design a recognition system based on the fusion of these distance parameters. Fig. 1 shows a person with its *VBB* and *PAB* boxes and the distances mentioned above for two postures: a sitting posture and the reference posture, which is the "Da Vinci posture".

3. BELIEF THEORY

The belief theory approach needs the definition of a universe Ω composed of N disjunctive hypothesis 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 cannot be classified in the following four static postures: standing (H_1), sitting (H_2), squatting (H_3) and lying (H_4). We add a hypothesis for the unknown posture class (H_0). Therefore we have Ω=$\{H_1$,H_2,H_3,$H_4\}$ and H_0. 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 is defined by:

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

with $\sum_{A \in 2^{\Omega}} m(A) = 1$. The measurements are the four independent distances D_i ($i = 1 \ldots 4$) presented in Section 2.

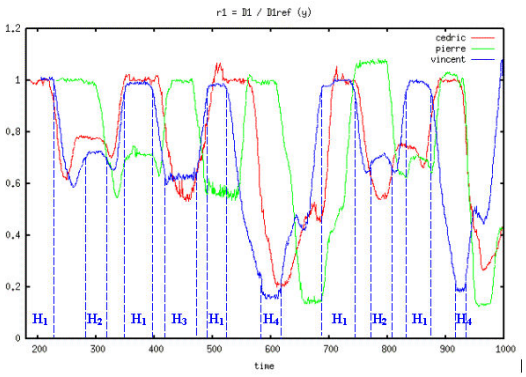

Fig. 2. r_1 variations for 3 different persons.

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 several persons in the same postures sequel: reference posture, sitting, standing, squatting, standing, lying, standing, sitting, standing and lying. The expected recognition results for the third person (Vincent) are shown at the bottom of Fig. 2 as the corresponding hypothesis label $H_{1\ldots4}$. As said before, we try to recognize the static postures and not the transitions between them.

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 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 reason is r_1 and r_2 vary in the same way, so do r_3 and r_4. The difference is that a measure in each pair is based on the face localization and not the other. All the thresholds of Fig. 3 were obtained after a human expertise over a training set of six different video sequences (see Section 4). The thresholds are different for each r_i.

3.2. 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 m_{r_i} and m_{r_j} is defined as follows:

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

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

Where A, B and C are subsets of 2^{Ω}. The orthogonal sum

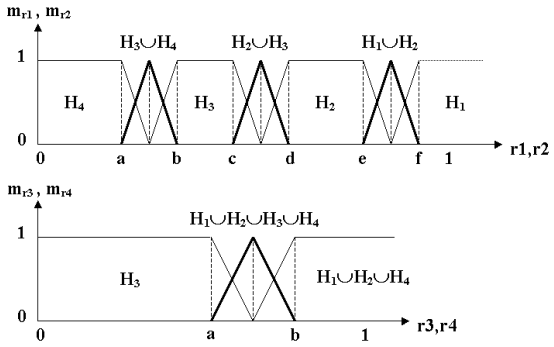

Fig. 3. Models m_{r_1}, m_{r_2} (top), m_{r_3}, m_{r_4} (bottom). H_i defines recognized posture(s).

Table 1. Combination table for the orthogonal sum of two distributions ($\emptyset$: 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

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

$$m_{r_{12}}(H_2 \cup H_3) = 0.8 \qquad m_{r_{34}}(H_3) = 0.9$$
$$m_{r_{12}}(H_2) = 0.2 \qquad m_{r_{34}}(H_1 \cup H_2 \cup H_3 \cup H_4) = 0.1$$

Their orthogonal sum gives the Table 1 combination table. The belief mass of each resulting subset is:

$$m_{r_{1234}}(H_3) = 0.72 \qquad m_{r_{1234}}(H_2 \cup H_3) = 0.08$$
$$m_{r_{1234}}(\emptyset) = 0.18 \qquad m_{r_{1234}}(H_2) = 0.02$$

The **conflict** is defined by the case of the empty intersection ($\emptyset$) of two subsets. It happens when the parameters are too different to obtain a coherent posture. For instance, grouping H_1, H_2 and H_4 (bottom model) prevents conflicts of being sit with a raised hand.

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 m, here $m_{r_{1234}}$, there is a choice to make between the different hypothesis H_i and their possible combinations. The choice has to be made with respect to the resulting belief mass distribution. Generally, 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. There are usual criteria used to make a decision:

$$\begin{aligned} \text{belief mass: } Crit(A) &= m(A) \\ \text{belief: } Crit(A) &= Bel(A) = \sum_{B \in 2^\Omega, B \subset A} m(B) \\ \text{plausibility: } Crit(A) &= Pl(A) = \sum_{B \in 2^\Omega, A \cap B \neq \emptyset} m(B) \end{aligned}$$

We chose the first criterion because models are quite simple. The accepted hypothesis is the **maximum belief mass singleton** because we look for a single posture. For the example of subsection 3.2, the decision is therefore H_3, which seems logical. If the conflict is maximum or if there are no singletons, we choose the hypothesis H_0, which means the posture is unknown.

3.4. 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 was: $H_0 = 0$, $H_1 = 1$, $H_2 = 2$, $H_3 = 4$ and $H_4 = 8$. Then the conjunction code for two combinations of hypothesis 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: $H_1 \cap H_2 = 01 \cap 10 = 00 = H_0$.

4. RESULTS

4.1. 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 video sequences resolution to $320 * 240$.

4.2. Training stage:

In this stage, six different persons have been filmed in the same ten successive postures. They were asked to be in "standard" postures, in front of the camera. Recognition rates are available on Table 2 confusion matrix. Columns show the real posture and lines the system recognized postures. As the models thresholds were determined by expertise over these training sequences, results are very good, except for squatting. The reason is everybody don't squat the same way, hands on knees or touching ground, back bent or straight etc. The average recognition rate is **88.2%**.

Table 2. Confusion matrix (training stage).

Syst\H	H_1	H_2	H_3	H_4
H_0	0%	11.1%	30.5%	5.5%
H_1	**100%**	0%	0%	0%
H_2	0%	**88.9%**	0%	0%
H_3	0%	0%	**69.5%**	0%
H_4	0%	0%	0%	**94.5%**

Table 3. Confusion matrix (testing stage).

Syst\H	H_1	H_2	H_3	H_4
H_0	0.3%	16.5%	31.7%	0%
H_1	**99.7%**	0%	0%	0%
H_2	0%	**79.8%**	24.7%	0%
H_3	0%	3.7%	**43.6%**	0%
H_4	0%	0%	0%	**100%**

4.3. Testing stage:

For the testing stage, six other persons have been filmed, in seven different successive postures. This time, they were "free", i.e. move the arms, sit sideways etc. Table 3 shows recognition rates in the confusion matrix. There are more recognition errors but the results shows a good recognition rate in general. 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 when people move their arms and raise them over their head. Most of the unknown postures recognized by the system are conflicts, showing that the models do not take into account all configurations of the "free" postures. The system detects a recognition problem instead of making a wrong choice because the distances ratio measurements are giving contradictory results. The average recognition rate for this stage is **80.8%**.

5. 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 parameters. 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. For instance, if a person moves far away from the camera, the system could be mistaken by recognizing this person as being sit because the person's segmentation mask is compact and not so elongated whereas he/she is in fact standing. One solution could be 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 parameters temporal evolution. This should greatly improve the recognition results. To justify this positive 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 real-time recognition of dynamic postures and actions recognition like standing up, sitting down etc.

6. REFERENCES

[1] 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] 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.