SILHOUETTE-BASED PROBABILISTIC 2D HUMAN MOTION ESTIMATION FOR REAL-TIME APPLICATIONS

Pedro Correa[1], *Jacek Czyz*[1], *Toshiyuki Umeda*[1], *Ferran Marqués*[2], *Xavier Marichal*[3], *Benoit Macq*[1]

[1] Communications Laboratory, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium
[2] Image Processing Group, Technical University of Catalonia (UPC) Barcelona, Spain
[3] Alterface SA, 1348 Louvain la Neuve (Belgium)

ABSTRACT

This paper presents a novel technique for 2D human motion estimation using a single non calibrated camera. The user's five crucial human features (head, hands and feet) are extracted, labeled and tracked, after silhouette segmentation. The crucial points candidates are defined as the local maxima of the geodesic distance with respect to the center of gravity of the actor region (silhouette) following the silhouette boundary. Selected crucial points are then classified as head, hands or feet using a probabilistic approach weighted by a prior human model. The system can run at 50Hz paces on standard Personal Computers.

1. INTRODUCTION

Currently, the development of human-computer interfaces, which enable a more natural communication mode between human beings and computers, is a very active area of research. A recent survey by Gavrila [1] classifies the research in this field in 3 categories: 2D approaches without explicit shape models, 2D approaches with explicit shape models and 3D approaches. This work falls into the second category. 2D systems do not require special equipment (such as stereo cameras [2]) or 3D articulated human model [3], and can therefore perform faster and less expensively than the 3D systems, while proving to be very useful and sufficient in a vast scope of different upraising areas, like virtual reality (interactive virtual worlds, teleconferencing), advanced user interfaces (gesture driven control) and motion analysis (gait-based biometrics, avatar animation). Technically speaking, this work is related to other 2D human body-tracking methods such as W4 [4], Pfinder [5] and Fujiyoshi's work in pedestrian detection [6]. Nevertheless, even if this above-mentioned articles have remarkable results in their goals, which is to track and/or monitor whole body motion, they do not contemplate interaction with separate body parts - and thus feature extraction - as a priority. W4 has impressive human surveillance capabilities but a fairly coarse body part detection and tracking, and Fujiyoshi's work only needs to approximately detect head and feet in order to achieve its pedestrian detection goal. Pfinder uses strong flesh-colored color priors. However, skin color is very environment dependent and may lead in some applications to unstable results [7]. Moreover color can also be unavailable as with infra-red lighting and for instance unusable to detect the feet. The system presented in this paper tracks all five human extremities (head, feet and hands) that overall define a specific human gesture using pure silhouette analysis, without relying on any kind of skin color detection or skin color tracking. As they are the minimal number of points that can be used in order to characterize human gestures [8], we will henceforth refer to these features as *crucial points* (CP). They are detected using geodesic distance maps computed on the body silhouette and subsequently labeled and tracked using a probabilistic approach weighted by an adaptive 2D human model. Results on real image sequences show a 94% average accuracy rate in crucial point detection and crucial point labeling and tracking. Please visit the following link in order to find test sequences and their results: `http://www.tele.ucl.ac.be/˜pedro/icip2005/`.

The paper is organized as follows. In the next section, we present an overview of the proposed algorithm. In Section 3 and 4 details of the crucial point detection and labelling are given, respectively. Section 5 is devoted to experiments and performance quantification on video sequences. Conclusions are given in Section 6.

2. ALGORITHM OVERVIEW

In order to extract the exact 2D posture of the human subject, called the *actor* in the sequel, we use one single non-calibrated camera. This way, the positions of the, at most, five crucial points (depending on the self-occlusions) are obtained. The actor silhouette or region is extracted using a real-time segmentation technique [9], which has very similar operating specifications as Pfinder's or other analogous advanced silhouette-based algorithms. As typically assumed for this kind of applications, we expect the scene

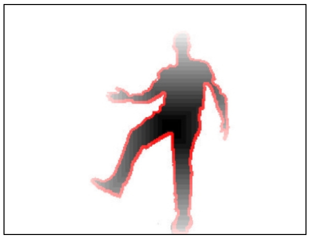

Fig. 1. Geodesic distance map for a given silhouette.

to be significantly less dynamic than the user and having overall changes that are small or gradual. The segmentation is achieved using a Walsh-Hadamard transform on blocks of 4x4 pixels. First, the module calculates the Walsh-Hadamard transform of the background image. Afterwards, the module compares the values of the Walsh-Hadamard transform of both the current and the background images. When the rate of change is higher than a threshold, this module classifies the area as foreground.

Once the silhouette is segmented, the posture estimation algorithm has two main steps:

Crucial point candidate selection: the algorithm extracts from the silhouette the points that are candidates to be crucial points. This first step is based on the analysis of the geodesic distance [10] between the silhouette and its center of gravity (CoG).

Crucial points labeling: once the crucial point candidates have been selected, we need to find out to which human features they correspond. We perform the labelling in two steps. In the first step (tracking), crucial points that were already labelled in the previous frame are matched with candidates crucial points. In the second step (detection), we assign to crucial point candidates labels that were not assigned during the first step.

3. CRUCIAL POINT CANDIDATE SELECTION

We define the crucial points as the five most prominent human features in the actor region. This notion of prominence can be translated in terms of distance from the CoG to the silhouette. We propose a robust method to search for crucial points based on the extraction of those points of the silhouette which represent local maxima of the geodesic distance with respect to the CoG. The CoG of the actor region is estimated by computing a weighted average of the region pixels [10]. The geodesic distance from the center of gravity C to any point x in the actor region A is defined as

$$d_A(C,x) = n \Leftrightarrow x \in \ \delta^n_A(C) \ \text{ and } \ x \notin \ \delta^{n-1}_A(C)$$

where the geodesic dilation of C using a structuring element of size n defined within A can be expressed as:

$$\delta^n_A(C) = \overbrace{\delta(...(\delta(\delta(C) \cap A) \cap A)...) \cap A}^{n \ \text{times}}$$

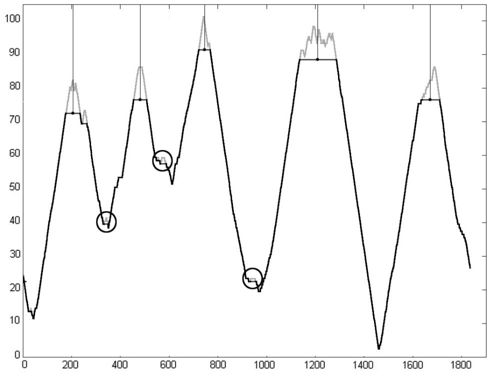

Fig. 2. Geodesic distance from center of gravity to silhouette border points. Local maxima correspond to prominent human features (crucial point candidates).

where δ is the morphological dilation. A one-dimensional function is computed representing the geodesic distance function from the CoG to the silhouette border points. This function contains the local maxima associated to the crucial points as well as other noisy local maxima. The latest ones are removed by operating a peak thresholding. Figure 1 presents the geodesic distance map from the CoG and the silhouette border (lighter border). In Figure 2, the function representing the geodesic distance evaluated on the silhouette is shown before (gray) and after (black) the threshloding. The removed local maxima are marked by circles and the selected local maxima are marked with vertical lines. This threshold is the factor with which we will be able to trade off noise removal against local maxima detection sensitivity. It is fixed once and for all depending on the camera and the working environment. The crucial point detection phase selects all local maxima still present after the thresholding step. In cases of self-occlusions, the number of detected crucial points can narrow down to two.

4. LABELLING OF CRUCIAL POINTS

The crucial point selection step outputs a list of candidate coordinates $z^{(i)}_t = (x, y)$ and associated *intensities* $I^{(i)}$ which are the *dynamics* of the geodesic distance local maxima [10]. The labelling algorithm classifies each couple $(z^{(i)}_t, I^{(i)})$ into one of the six classes: head, left foot, right foot, left hand, right hand and noise, denoted by $\Omega = \{h, lf, rf, lh, rh, n\}$. The algorithm works in two steps: (i) tracking existing crucial points and (ii) detecting new appearing crucial points which were not visible or not detected in the previous frame. In the tracking step, the temporal continuity of human motion permits to use the previous position of a given crucial point to aid the classification. To classify a candidate $z^{(i)}$ we use a Maximum a Posteriori (MAP) rule. We compute $P(\omega_\alpha|z_t, z_{t-1})$ for each $\omega_\alpha \in \Omega_T \cup \{n\}$ (Ω_T

is a subset of Ω of tracked points) and assign to the point the class that has maximum probability, i.e.

$$\omega^* = argmaxP(\omega_\alpha|z_t, z_{t-1}).$$

Using Bayes law, the a posteriori probability can be written as a product of three factors, i.e.

$$P(\omega_\alpha|z_t, z_{t-1}) = \frac{p(z_t, z_{t-1}|\omega_\alpha)P(\omega_\alpha)}{p(z_t, z_{t-1})} \quad (1)$$

$$\propto p(z_t|\omega_\alpha)p(z_t|z_{t-1}, \omega_\alpha)P(\omega_\alpha), \quad (2)$$

the normalising factor $p(z_t, z_{t-1})$ can be discarded since it does not influence the classification. The first factor $p(z_t|\omega_\alpha)$ represents the prior knowledge available on the presence of a crucial point (of class ω_α) at location z_t. This term is computed using a prior probability map described in Section 4.1. We suppose that the crucial points kinematics, represented by the second factor, can be modelled as a Gauss-Markov random sequence, i.e.

$$p(z_t|z_{t-1}, \omega_\alpha) = \mathcal{N}(z_t; z_{t-1}, S_\alpha), \quad (3)$$

where the covariance S_α is chosen diagonal. Note that instead of using the crucial point position in the previous frame as the mean of the gaussian, we could use a predicted position $\hat{z}_t$ using a first order dynamic system. As for the third factor $P(\omega_\alpha)$, it reflects the prior knowledge on class ω_α. In our case the classes head and feet have a higher $P(\omega_\alpha)$ than hands since hands are much more often occluded than head and feet.

In the detection step, we try to find new crucial points, if any, that were occluded or not detected before. We thus classify the remaining candidate points in the remaining classes $\Omega \setminus \Omega_T \cup \{n\}$ applying the same technique but using the probability map, the prior and the intensity of the candidate crucial points:

$$P(\omega_\alpha|I_t, z_t) \propto p(z_t|\omega_\alpha)P(\omega_\alpha)p(I_t|\omega_\alpha) \quad (4)$$

where $p(I_t|\omega_\alpha)$ is modelled by uniform density with different parameters for noise and crucial point classes.

Finally, the system does not need any kind of forced initialization, since for the first frames of a sequence, the system simply works in pure detection mode until reliable crucial points are found.

4.1. Prior probability maps

In this section we describe the probability maps used for computing the factor $p(z_t|\omega_\alpha)$ in equation (2). The map reflects our prior knowledge about possible location z_t of the crucial point ω_α. For each frame, a generic map is adapted to the observed silhouette using its height, center of gravity and torso angle. To estimate the torso angle, the silhouette is eroded using a large structuring element so that only the torso area remains. The torso angle is defined as the angle of the largest axis of the torso area.

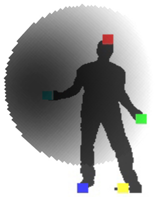

Fig. 3. prior probability map for crucial point label $\omega_\alpha = rh$ (right hand).

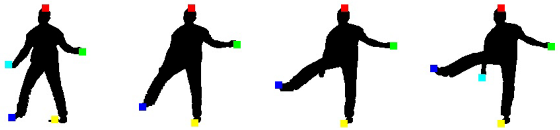

Fig. 4. Typical results on 4 frames. On the third frame, the hand is already visible but not yet detected by the algorithm.

Figure 3 shows the prior probability map for crucial point rh (right hand). Although better results are expected if the map is learned using real human motion, we experienced that the map does not need to be extremely accurate to achieve good results. Note that the map can be easily adapted for analysing different motion types. This enables a generalisation of the method to other walking creatures and other applications.

5. EXPERIMENTAL RESULTS

For performance quantification purposes we analyze the same sequence using a real segmentation and a perfect segmentation performed manually. This allows to separate pure tracking errors from errors induced by the segmentation. The test sequence is 380 frames long and displays a vast scope of human gestures: jumps, walks, hand gesticulation in all directions and hand and feet inversions and occlusions. It also represents a good sample of realistic working conditions. This particular sequence and some other real and synthetic ones and their results are available at:
`http://www.tele.ucl.ac.be/~pedro/icip2005/`

We assume that the region of the hand extends from the tip of the fingers to the wrist, and the feet from the ankle to the tip of the toes. An error is counted each time a label is detected outside these regions, or not detected although it is distinguishable by visual inspection. In figure 4 for example, the right hand is not detected in the third frame, yet being slightly visible: it is thus counted as a missed detection error. We thus divide the possible errors in two different sub-classes: erroneous label and missed detection. These errors are reported in Table 1. The sum of missed detection and label errors divided by the total number of frames gives

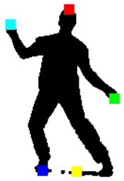

Fig. 5. Segmentation errors due to residual shadows lead to small foot location bias.

the error rate listed in the table. The average error rate has a value of **5.5%** (**1.9%** with perfect segmentation). Table 1 shows also error rates obtained using the same sequence but manually segmented. The higher error rates found during the feet tracking using automatic segmentation are due to residual shadows segmented below the user. In those cases feet are labeled very near to the exact foot location, but still outside the region we defined previously (see Figure 5). This also explains the dramatic drop in error rates concerning the feet when comparing real segmentation to perfect segmentation. Note that the errors are more severe for the left foot since it creates stronger cast shadows in this sequence.

	Left Hand	Right Hand	Head	Left Foot	Right Foot
Automatic Segmentation					
Label Error	4	6	1	47	22
Missed Dectect.	8	4	3	2	4
Error Rate (%)	3.3	2.7	1.0	13.3	7.0
Manually Corrected Segmentation					
Label Error	4	6	1	1	2
Missed Detect.	8	4	3	2	4
Error Rate (%)	3.3	2.7	1.0	0.8	1.6

Table 1. Performance on the test sequence.

The most time consuming part of the algorithm is the geodesic distance map creation, the other parts are very light from a computational point of view. Nevertheless, a recently proposed approximation [11] of the geodesic distance computation has been optimized and runs at 150 Hz for image segmentation and body analysis on a standard PC (Pentium IV CPU at 1.5GHz and 256MB of memory). The overall algorithm is able to run at 50 Hz frequencies under the same conditions.

6. CONCLUSIONS

A new technique for robust real-time 2D human posture estimation has been proposed. The method relies on the analysis of geodesic distance maps computed on the actor region and MAP labelling supported by a statistical human model. The method is very stable and fast compared to other techniques. Currently, we are working in the possibility of increasing its detection sensibility during self-occlusions using a skin-detection technique as a back up.

7. ACKNOWLEDGMENTS

This work has been partially supported by the Walloon Region under the FIRST EUROPE Objectif 3 grant EPH3310 300R0132 /215069, the EU NoE SIMILAR and the grant TIC2001-0996 of the Spanish Government.

8. REFERENCES

[1] D.M. Gavrila, “The visual analysis of human movement: A survey,” *Computer Vision and Image understanding*, vol. 73, no. 1, pp. 82–98, January 1999.

[2] D. Demirdian, and T. Darrell, “3-D Articulated Pose Tracking for Untethered Diectic Reference,” *Int. Conf. Multimodal Interfaces*, Pittsburg, 2002.

[3] P. Fua, A. Gruen, N. D’Apuzzo, and R. Plankers, “Markerless full body shape and motion capture from video sequences,” *Int. Arch. of Photogrammetry and Remote Sensing*, vol. 34, no. 5, pp. 256–261, 2002.

[4] I. Haritaoglu, D. Harwood, and L.S. Davis, “W4: Real-time surveillance of people and their activities,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 809–830, August 2000.

[5] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “Pfinder: Real-time tracking of the human body,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 780–785, July 1997.

[6] H. Fujiyoshi and A. Lipton, “Real-time human motion analysis by image skeletonization,” in *IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 15–21. October 1998.

[7] M. Soriano, B. Martinkauppi, S. Huovinen, and M. Laaksonen, “Skin detection in video under changing illumination conditions,” in *Proc.of the 15th Int. Conf. on Pattern Recognition*. September 2000.

[8] G. Johansson, “Visual perception for biological motion and a model for its analysis,” *Perception and psychophysics*, vol. 14, pp. 201–211, February 1973.

[9] X. Marichal, B. Macq, D. Douxchamps, and T. Umeda, “Real-time segmentation of video objects for mixed-reality interactive applications,” in *SPIE Visual Communication and Image Processing Conference*, pp. 41–50. Hilton Head, SC, July 2003.

[10] Soille P., *Morphological Image Analysis, Principles and Applications, Second Edition*, Springer-Verlag, 2003.

[11] T. Umeda, P. Correa, F. Marqués, and X. Marichal, “A real-time body analysis for mixed reality application,” in *Korea-Japan Joint Workshop on Frontiers of Computer Vision, FCV-2004*. February 2004.