Unsupervised Human Members Tracking Based on an Silhouette Detection and Analysis Scheme

Costas Panagiotakis and Anastasios Doulamis

Abstract In this paper, an unsupervised, automatic video human members(human body and parts) tracking algorithm is proposed based on a propabilictic human silhouette detection method and a geometric human silhouette analysis algorithm. First, the human silhouette is estimated using the following scheme : 1) an face detection, and 2) a human body detection based on biometric 2-D templates. Next, the human members are recognized and 18 major points, located on the human body and parts, are detected using a 2-D biometric model. The proposed method is executed automatically in each frame of the video sequence.

1 Introduction

Human motion analysis has many applications in many areas, such as analysis of athletic events, surveillance, content-based image storage and retrieval. The main scientific challenges in human motion analysis are to detect, track and identify people and to recognize the human activity [1]. The detection and tracking algorithms are challenged by occluding and fast/complicated moving objects, as well as illumination changes. A combination of human shape-motion features estimation, silhouette analysis, skin color detection, template matching, 2–D/3–D human modeling, background modeling have been used on human detection and tracking systems. There are model based approaches and systems using Shape-From-Silhouette methods to detect and track the human in 2D [2] or 3D space [3].

Several works have been proposed recently in the literature for detecting video actions and activities [2]. In [4], a Hidden Markov models are used for identifying

Costas Panagiotakis
Department of Computer Science, University of Crete, Heraklion, Greece, e-mail: cpanag@csd.uoc.gr

Anastasios Doulamis
Technical University of Crete, Greece, e-mail: adoulam@cs.ntua.gr

human activities in video streams. In addition, in [5] human activity is identified from a video sequence. The activity is represented as a set of pose and velocity vectors for the major body parts (hands, legs, and torso) and then stored in a set of multidimensional hash tables. On the contrary, in [6], stochastic algorithms are exploited to detect visual activities and interactions in a video sequence.

In this paper, a novel approach is presented for automatic human detection and tracking in video sequences. The proposed architecture consists of two main modules. The first refers to the automatic identification of humans regardless of their motion actions, background complexity and possible background movement. This part is based on biometric feature of human color. The second module aims at automatically extracting points of interest from the human based on 2-D biometric human model. The points refer to salient parts of persons, such as the head, shoulders, etc. In this way, we are able to identify complex human activities, by considering the point motion in time. For example, we are able to detect particular gestures of humans, the direction of their movement, the actions of the head, e.g., nodding, and velocity of human movement, etc.

The rest of the paper is organized as follows. Section 2 presents the proposed silhouette estimation algorithm. Section 3 describes the human silhouette analysis method. Finally, Sections 4, 5 provide experimental results and the discussion.

2 Human Content Identification

Various methods and algorithms have been proposed in the literature over the years for human face detection, ranging from edge map projections to recent techniques using generalized symmetric operators [7]. The eigentemplate approach to the detection of facial features has been proposed in [8], while in [9] the face pixels are localized by modeling the human face with an elliptic shape. In our approach, the two-chrominance components of pixels are used for performing the human face detection task efficiently, while simultaneously exploiting information available in the bit stream of MPEG-coded images. This is due to the fact that the distribution of the two-chrominance components, corresponding to a human face, are located in a very small region of the color space as has been shown in [10]. Thus, blocks of a color image x, whose respective chrominance values are located at this small region, can be considered as face blocks. On the contrary, blocks of chrominance with values located far from this region correspond to non-face blocks.

Let us denote by $q(B_i) = [u(B_i)v(B_i)]^T$ a 2-dimensional vector containing the average chrominance components, $u(B_i)v(B_i)$, for the B_i block. Then, the histogram of the chrominance values, corresponding to the face area, is modeled by a Gaussian probability density function (pdf). Therefore, the probability of a block, say B_j, belonging to the face class, say Ω_f, is given by the following equation

$$P(q(B_j)|\Omega_f) = \frac{e^{-\frac{1}{2}(q(B_i)-\mu_f)^T \cdot S_f^{-1} \cdot (q(B_i)-\mu_f)}}{(2\pi)^{\frac{N}{2}} |S_f|^{\frac{1}{2}}} \qquad (1)$$

where μ_f and S_f are the mean vector and variance matrix of the pdf respectively. The parameters of 1 can be estimated based on several training data of face images and using the maximum likelihood algorithm. Equation 1 indicates that an image block B_i belongs to the face area, if the respective probability of its chrominance values, $P(q(B_i)|\Omega_f)$ is high. Instead, blocks with a low probability $P(q(B_i)|\Omega_f)$ are classified as non face blocks. In our case, a confidence interval of 80% has been used to discriminate face and non face blocks. Therefore, a binary mask M is formed, with size $\frac{N_1}{8}x\frac{N_2}{8}$ pixels; a pixel unit with value equal to one indicates a face block, while a zero value indicates a non face one.

This method does not, however, exploit any geometric information about the human face. Thus, it is possible that some non face blocks are classified as face ones. This is, for example, the case of blocks that have similar chrominance properties to ones belonging to face regions, e.g., human hands. For this reason, an iterative technique is applied to the binary mask M in order to localize the segment that corresponds to the face region. First, the morphological erosion operator is applied to image M, using a small rectangular structuring element; then, the number of non connected objects in the filtered mask is computed. In case that the number of objects is greater than one, a new morphological filtering is applied using, however, a greater structuring element. This procedure iterates until the number of objects gets equal to one. Then, the segment of M, which overlaps to the segment of the final filtered mask, is considered as a face region. In this case, a binary mask, say M_f, is formed, in which pixels with value equal to one correspond to the face segment, while zero values indicate the other areas.

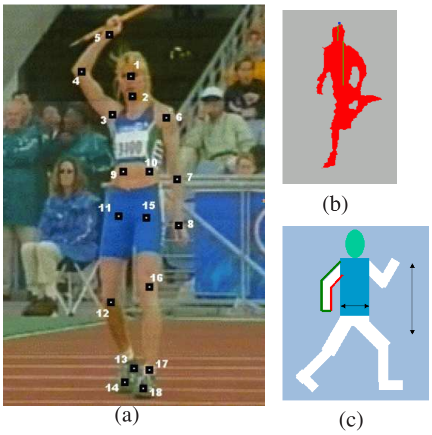
2.1 Human Body Detection

Human body detection is next performed, exploiting information provided by the previous face detection module. In particular, the human body is localized using a probabilistic model, the parameters of which are estimated according to the center, height and width of the face region, denoted as $c_f = [c_x\ c_y]^T$, d_f and h_f respectively. Let us also denote by $r(B_i) = [r_x(B_i)\ r_y(B_i)]^T$ the distance between the ith block, B_i, and the origin, with $r_x(B_i)$ and $r_y(B_i)$ the respective x and y coordinates.

Since humans are usually located in standing position, a square rectangular is adopted in our case for modeling human body. We assume independence from the x and y location. Thus, block B_i belongs to the human body class, say Ω_b, if

$$P(r(B_i)|\Omega_b) = A \cdot [(1 - \frac{r_x(B_i) - \mu_x}{w_x}), (1 - \frac{r_y(B_i) - \mu_y}{w_y})] \tag{2}$$

where μ_x, μ_y express the parameters of the human body location model; these are calculated based on information derived from the face detection task, taking into account the relationship between human face and body. In our simulations, the



(a) (b) (c)

Fig. 1 **(a)** The 18 major human points. **(b)** An example of successful execution of the end of head (blue point) localization method. **(c)** The left arm can be distinguished from the main body because the proportion of red boundary pixels is high.

parameters in 2 are estimated with respect to the face region as follows $\mu_x = c_x$, $\mu_y = c_y + \frac{h_f}{2}$, $\sigma_x = d_f$, $\sigma_y = \frac{h_f}{2}$.

Similarly, A,w_x and w_y are appropriate parameters that control the decay of the probability.

The human face and body detection modules provide an initial estimation of the foreground object. In particular, all blocks that have been classified either to the face or body classes are included in the initial estimate of the foreground object. Similarly, a background set, is created containing blocks of the image which are classified with high confidence to the background class.

3 Human Silhouette Analysis

3.1 Human Members Recognition

In this section we examine the human body members recognition method. The human body is divided into the following six members: head, main body, left leg, right leg, left arm and right arm (Figure 2(c)). The human silhouette pixels will be classified to one of the above members. The human parts are detected based on their geometric/biometric information. The member recognition algorithm is sequential. The more “visible” members are computed first in order to decrease the search space of others.

First, the major human body axis is determined using central moments. The silhouette is rotated according to the major axis. The rotation center is the mass center

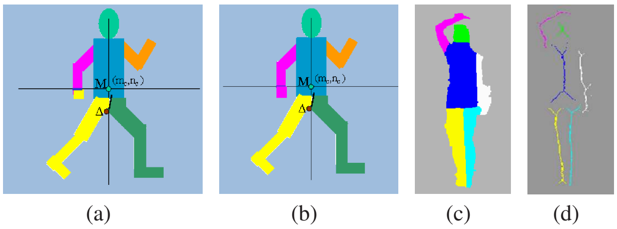


Fig. 2 **(a)** The initial approximation. **(b)** The final segmentation result. **(c)** The human members estimation. **(d)** The human members' skeletons.

of the human. Now, the human is located vertically. Next, we estimate the end point of the head (X_d, Y_d) using the following iterative method (Figure 1(b)). The (X_d, Y_d) point is initialized as the human body mass center (m_c, n_c) and it is changed dynamically by getting the mean of human points that belongs to the current line. So, the above method can be characterized as dynamic mean method.

Dynamic Mean Method

```
$X_d = m_c, \quad Y_d = n_c$
repeat {
    $X_d = X_d - 1, \quad Y_d = E_{(X_d, y) \in Man}(y)$
} until $(X_d - 1, Y_d) \in Background$
```

The head region will be placed in a rectangle defined by the point (X_d, Y_d) as the end of the head. The maximum height and width of the head are proportional to the human height. The down limits of the head region, are determined by the first local minimum of the left and right horizontal silhouette projections.

The main body region is computed using an iterative algorithm similar to the end point of head estimation method. The maximum height and width of the main body are proportional to the human height. Using the human boundary, it can be determined if the arms can be distinguished from the main body (visible arm) and if the legs are distinguished. The rule is the following: if the proportion of the boundary pixels whose the closest background pixel is on the right, computed in an area where is located the left arm, exceeds a threshold, then the left arm can be distinguished from the main body. The above value is the ratio between the red pixels and the red plus green pixels of Figure 1(c). This knowledge helps in definition of the main body limits.

The legs and arms regions estimation can be done in the same time. An initial approximation (Figure 2(a)) is computed using the mass center of the human and the border point D that discriminates the left and right leg. It is possible that a regions than six, that is the number of the human members, will be created. For these cases, we determine first which are the fault regions using a criterion based on the surface and the mass center of the regions. Next, we join the faults regions to the regions which have the most common boundary points with the faults regions. In the Figure 2(b), the false left leg region is correctly classified to left arm member.

3.2 Major Human Points Estimation

In the second stage, the 18 major human points are estimated using the human members segmentation. This method is based on a geometric/biometric 2-D model for each major human point. The method uses the skeleton of each human member (Figure 2(d)), as the joint points belong in the skeletons. The skeleton is defined as the set of points whose distance from the nearest boundary is locally maximum. The 18 major human points are computed sequentially. The easier defined points are computed first in order to decrease the search space of others.

First, the center of the head is computed as the mass center of the head region. The neck point is defined as the mean of the boundary points between head and main body region. The two shoulders points are computed by minimizing an appropriate function F. The function domain is an isosceles triangle whose vertex is the neck point and its base vertices are the two shoulder points. The function is minimized when the triangle base is maximized and the triangle height is minimized at the same time.

The 18 major points formulations are defined in Figure 1(a) under the proposed biometric 2-D model. The points (9), (10) of the main body are computed using the main body height and width. The points (11), (15) of the main body are defined by the mean of boundary between main body region and left or right leg region respectively.

Concerning the legs' points, the ankle point A is computed first. We compute the farthest point B of skeleton points from point (9) using one line segment that should belongs to silhouette. The ankle point is defined as the farthest point, of the not visited skeleton points from B using one line segment. The knee point K is estimated by minimizing the function $G(X)$ which is defined by the following equation. The constant 0.2 of equation 3 has been estimated using our experimental dataset. Let F be the point (9) of the main body. Let the function $d(X,AF)$ be the minimum distance of point X from the line segment AF.

$$G(X) = (|XF| - |XA|)^2 - 0.2 \cdot d^2(X,AF) \tag{3}$$

If the point X is located close to the middle of AF and close to the knee angle at the same time, then the proposed function $G(.)$ will be minimized. Finally, the end of leg point E is computed using the knee and ankle points. The E point is defined as the skeleton point close to ankle point, whose distance from the knee point is maximum. In each arm, we have to compute two points, the elbow point and the end of arm point which are estimated similarly with the knee and ankle point, respectively.

4 Results

The proposed algorithm have been tested in several sequences. The silhouette estimation method (first stage) gives in most of the frames accurate results. Concerning

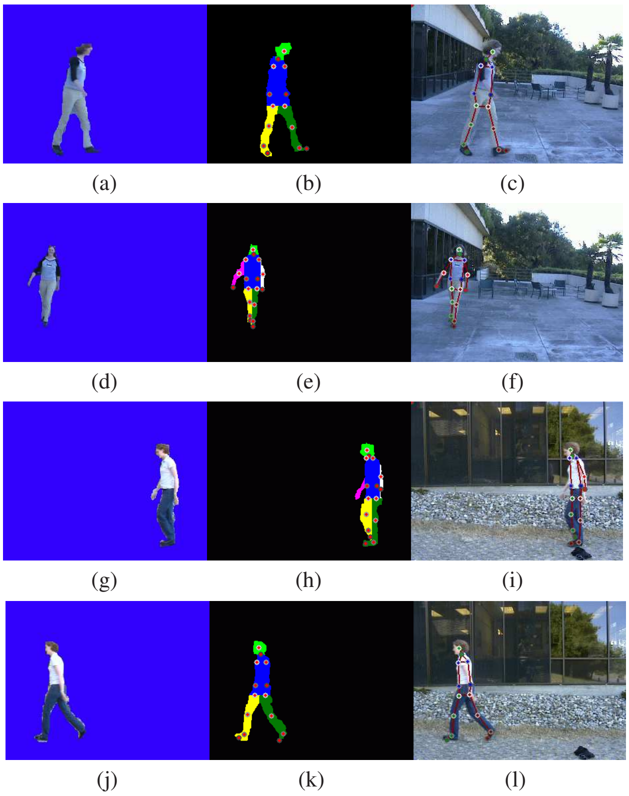


Fig. 3 Results of the three stages. The estimated silhouettes (first column), the human members recognition (second column) and the 18 major human points computation (third column).

the next two stages, the human body members are successfully recognized in the well estimated silhouettes with high accuracy. The mean error in head, main body major points estimation is about 2% of the total human height, while the arms - legs major points are estimated with less than 200% of the previous error. The above errors are related with the quality of the estimated silhouette, when the accuracy of the estimated silhouette is high the mean error in the estimation of each major human point is less than 1% of the total human height. In Figure 3, some results of the three stages are shown. The complexity of the total algorithm is $O(N)$[1].

[1] N = # pixels of the human. If we did not use skeletons we would have a complexity of $O(N\sqrt{N})$.

5 Conclusion

In this paper, an unsupervised, automatic human members and 18 major human points tracking algorithm is proposed using an adaptable - extended face detection based method and a geometric human silhouette analysis algorithm. The adaptive behavior of the first stage is very important in such dynamically changing environments, where object properties frequently vary through time. The silhouette analysis algorithm is color independent and it detects the major human points without tracking them based on 2-D geometric human model.

In order to decrease the computation cost, the proposed method locates the major points sequentially, where the location of one feature influences the location of the rest. Sometimes, such methods can produce totally erroneous results in the case that a failure occurs in the starting stages. We decrease this probability by starting from the most "visible" parts and well defined points. Moreover, if in some frame the algorithm fails (in some points/parts), the system will not lose his stability, since in the next frame human detection (not tracking) will be performed.

An extension of the proposed methodology may include the estimation of static (chromatic-biometric) features of the human members and the human-activity recognition (walking, sitting, running, etc). Security system and statistics analysis of human motion systems could be based on our method.

References

1. J. Aggarwal and S. Park, "Human motion: Modeling and recognition of actions and interactions," in *3DPVT04*, 2004, pp. 640–647.
2. C. Panagiotakis, E. Ramasso, G. Tziritas, M. Rombaut, and D. Pellerin, "Shape-motion based athlete tracking for multilevel action recognition," in *Proc. of AMDO 2006*, 2006, pp. 385–394.
3. K.M. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette across time: Part ii: Applications to human modeling and markerless motion tracking," *Int. Journal of Computer Vision*, vol. 63, no. 3, pp. 225–245, 2005.
4. M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," *IEEE Trans. on PAMI*, vol. 22, no. 8, pp. 844–851, 2000.
5. J. Ben-Arie, Zhiqian Wang, P. Pandit, and S. Rajaram, "Human activity recognition using multidimensional indexing," *IEEE Trans. on PAMI*, vol. 24, no. 8, pp. 1091–1104, 2002.
6. Y.A. Ivanov and A.F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *IEEE Trans. on PAMI*, vol. 22, no. 8, pp. 852–872, 2000.
7. D. Reisfeld, H. Wolfson, and Y. Yeshurum, "Detection of interest points using symmetry," in *Proc. of Inter. Conf. Coputer Vision*, 1990.
8. B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *Pattern Analysis Machine Intelligent*, vol. 19, no. 3, pp. 696–710, 1996.
9. A. Eleftheriadis and A. Jacquin, "Automatic face location for model-assisted rate control in h.261 compatible coding of video," *Signal Processing Image Communication*, vol. 7, pp. 435–455, 1995.
10. H. Wang and Shih-Fu Chang, "A highly efficient system for automatic face region detection in mpeg video sequences," *IEEE Trans. on Circuits and Syst. for Video Technol,special issue on Multimedia Systems and Technologies*, 1997.