AUTOMATIC HUMAN BODY TRACKING AND MODELING FROM MONOCULAR VIDEO SEQUENCES

[1]Chih-Chang Yu, [2]Jenq-Neng Hwang, [1]Gang-Feng Ho and [3]Chaur-Heh Hsieh

[1]Department of Computer Science and Information Engineering
National Central University, Taiwan
[2]Department of Electrical Engineering, Box 352500
University of Washington, Seattle, WA 98195
[3]Department of Information Engineering
I-Shou University, Taiwan

ABSTRACT

In this paper we developed a system for automated human body tracking and modeling based on a monocular camera. In this system, eleven joint points including head, shoulder, hip, elbows, knees, hands and feet are extracted separately to build a 2D human body model. The head is extracted by analyzing negative minimum curvature (NMC) points on a parameterized silhouette. The torso, along with its angle and size, is determined by integrating multiple frame information with connectivity constraint. Hands and feet can be identified correctly based on a modified star skeleton approach and the nearest-neighbor tracking mechanism. The rest of joint points can also be located by taking advantage of the connectivity constraints. A successful construction of the proposed human body modeling will pave a critical foundation for further intelligent analysis in many applications, such as automated video surveillance system or systematic video understanding.

Index terms*:* Human body modeling, negative minimum points, intelligent analysis

1. INTRODUCTION

In this paper, we develop an automated human body extraction and behavior analysis system for video surveillance and video understanding/analysis applications. Usually a human body can be approximately represented by several parts such as head, torso, and limbs with some joints like shoulders, elbows, hips and knees. Recently Gavrila [1] provides a comprehensive survey and divides this research field into three categories: 2D approaches without shape models, 2D approaches with shape models and 3D model approaches. Research proposed by Viola et al. [2] belongs to the first category. They integrate image intensity information with motion information to detect walking pedestrians. One approach in the second category model a person based on a single 2D image [3] . The others are based on body part tracking techniques, either using one single body part [4] or the whole body [5].

The challenge on shape-based human modeling approach is that human limbs are often mis-detected due to the self-occlusion or occluded by other scene objects. A fast and simple approach to extract limbs end from silhouette is star skeletonization [6], which identifies the extremities on the contour boundary and connects them with centroid of gravity (CoG). Due to the self-occlusion problem, the extremities are not always five, not to mention some false positives are included and confuse the limbs end extraction.Furthermore, the star skeleton is not a good match for the real human model when they bend their limbs. Therefore, in this paper we propose an innovative approach to construct a more reliable body model from video. Our approach starts by finding two important joint points: shoulder and hip after locating the head based on curvature analysis. In the meanwhile, we integrate multiple frames information to identify the extremities as hands or feet. Moreover, we employ a notion called "connectivity energy" to locate the elbow and knee joint points. A complete human model can thus be effectively built for behavior recognition and analysis purpose.

The rest of the paper is organized as follows: Section 2 discusses our techniques for video object extraction, head and torso extraction, limbs tracking and association, and joint points locating to build a complete model. Section 3 shows the performance evaluation about our modeling mechanism followed by the Conclusions in Section 4.

2. AUTOMATIC HUMAN BODY MODELING SYSTEM

Figure 1 shows the system framework on our automatic human body parts extraction/tracking system which starts with a background subtraction to obtain the video object, a negative minimum curvature (NMC) based cuts generation, head and torso extraction, extremities extraction and limbs

tracking and modeling. Details of these modules are given in the following subsections.

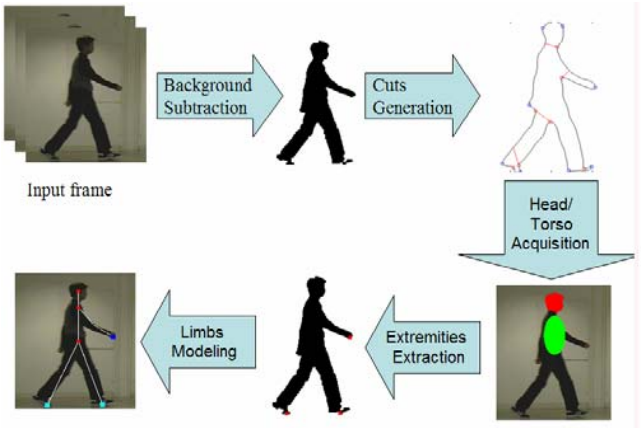

Figure 1: the flowchart of the system.

2.1. Video Object (VO) Extraction

Video objects (VOs) can be extracted by employing the background subtraction algorithm. Each image frame is subtracted from the background image to obtain the difference image. The fourth order moment of each pixel in the background difference image is calculated using the following equation [7]:

$$\mu_d^{(4)}(x,y) = \frac{1}{N_\eta} \sum_{(s,t)\in\eta(x,y)} (diff\_img(s,t) - \hat{m}_d)^4 \quad (1)$$

where $\eta(x,y)$ denotes a window of size N_η. Each pixel (x,y) in the temporal difference image is thresholded based on $\mu_d^{(4)}(x,y)$. Afterwards, some morphological operations (opening and closing) are performed to get rid of noises.

With a fixed camera and indoor environment, the background subtraction algorithm performs reasonably good segmentation after refinement by morphological post processing.

2.2. Human Body Model

A 2D human body model used in our experiments is shown in Figure 2. It is composed of 11 parts which are head, shoulder, hip, elbows, knees, hands and feet. How to decompose automatically a human body into these parts has become a challenging issue. Hoffman and Richards [8] propose an idea of transversality which is formulated as the following rule: Points of maximally concave extrema, or technically called negative minimum curvature (NMC) points, are good candidates for part-boundaries. Siddiqi and Kimia [9] propose two rules, Limbs and Necks, for fitting body parts into silhouette shapes. Considering all human body parts the head is the easiest one to locate and segment. In this paper we only perform the neck rule of NMC on head acquisition because the head and the torso can be treated as a near-rigid object thus we can usually get a cut to separate the head and the torso.

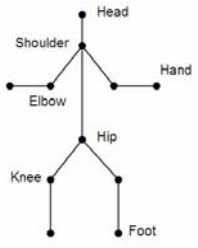

Figure 2: 2D human body model

2.3. Head and Torso Acquisition

After VO extraction the human silhouette is obtained (see Figure 3-b) and the NMC points can be obtained by computing the second order partial derivatives on the extracted contour boundary . Unfortunately the zigzag silhouette results due to imperfect VO segmentation contain plenty of false NMC points. Therefore we first use the cubic B-spline interpolation [10] to transform the discrete boundary data to continuous data (see Figure 3-c). The NMC points can be easily obtained by computing the second order partial derivatives on the parameterized silhouette (Figure 3-d). Then we employ the necks rule on these NMC points to create cuts through these NMC points and segment the VO into several sub-regions (Figure 3-e). We have to decide which sub-region is the head region of the first frame, which can then be used for the subsequent frame head tracking purpose. Among these sub-regions we compute the major and minor axis of each sub-region. Because the head is roughly in round shape, we choose the sub-region whose ratio of major and minor axis is closest to 1. Figure 3-f shows the results of the final extracted head region. Subregions of the rest of frames can be determined as the head region or not by using Kalman Filter tracking [11].

2.4. Torso Extraction

After obtaining the head region, the location and the length of the torso can be inferred by taking advantage of head coordinates. We use an intuitive definition called "connectivity" which means two points belong to the same region if you can make a straight line between them without intersecting any silhouette boundaries. A connectivity energy function is defined as follow:

$$E(i,j) = \begin{cases} D(i,j) & if \quad connectivity(i,j) = 1 \\ 0 & , \quad otherwise \end{cases} \quad (2)$$

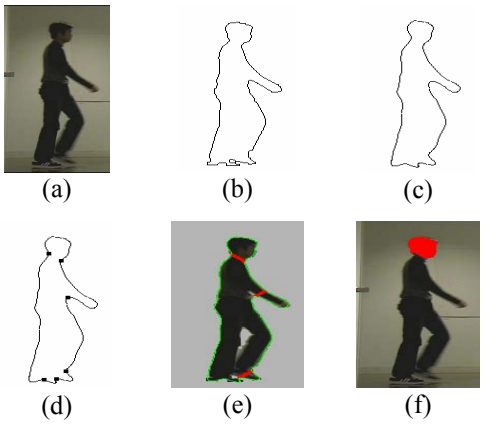

Figure 3: Examples of head acquisition results. (a) Original image (b) discrete contour (c) B-spline interpolation (d) NMC points (e) cuts which are employed necks rule on NMC points (f) head region.

where D is the Euclidean distance function. The connectivity information can tell us whether two points belong to the same body parts or not and this information is also applied on our joint points finding. Due to the kinematic constraint of degree of freedom (DOF), the motion of the head is limited in a small range. As shown in Figure 4-a, if we try to find the largest connectivity energy from the centroid of the head to the contour boundary, this line will most likely pass through the torso part. After obtaining this line we set the head centroid as the center, align (rotate) all VO images to make this line perpendicular to the ground plane, sum them together to produce a gesture map (see Figure 4-b) and a clean torso region can be obtained after appropriate thresholding (see Figure 4-c). Then we use an ellipse to fit this region and get the major axis and minor axis of the torso (see Figure 4-d).

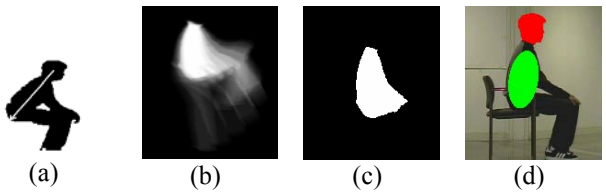

Figure 4: (a) Maximum connectivity energy from the head (b) Gesture map (c) The torso region (d) Results of extracted head and torso region

2.5 Limbs Classification

Based on the torso ellipse we can further infer two joint points, shoulder and hip, which are the foci of the torso ellipse. After extracting these two joint points we start to classify the extremities into two classes: hand and foot. There is no need to consider the head extremity because we have already extracted the head region in the previous step. The extremities can be found by using the approach in [6]. We use the shoulder point as the reference point rather than the CoG used in [6]. For the first frame the extremities are classified into two classes based on their distance to the shoulder. Then we use a nearest-neighbor tracking algorithm to associate the extremities between frames. The algorithm is stated as follows:

1. For a given extreme points p_k^i in frame i, compute the Euclidean distances between and all extremities points p_j^{i-1}, where j=1...n. If a nearest neighbor p_j^{i-1} which $D(p_k^i, p_j^{i-1})$ is below a threshold exists, assign p_k^i as the same class as p_j^{i-1}. D is the Euclidean distance.
2. If p_k^i is not close to any p_j^{i-1}, $j = 1...n$, treat p as a new extremity and assign it to one of the two class. If p is close enough to the shoulder, assign it to class "hand", otherwise assign it to class "foot".
3. If there are more than two extremities which are assigned to the same class, start from the point p_{far}^i which is farthest from shoulder/hip. If p_{far}^i is originally classified as “hand”, the system will assign this point to “foot” if the class “foot” does not have two extremities associated with it. Otherwise discard p_{far}^i. If p_{far}^i is originally classified as “foot”, discard p_{far}^i directly. Doing Step 3 recursively until the number of points in this class is 2. (Because a human can only has 2 hands and 2 feet at most.)
4. Repeat until all frames are processed.

2.6 Elbow and Knee Positioning

After the limbs classification step in Section 2.5, the hand extremities should connect to the shoulder and the foot extremities should connect to the hip through two more joint points: elbow and knee. Human limbs can be defined as two sets: hand-elbow-shoulder and foot-knee-hip. However, most human’s motions are so complicated that these points are not always in a row. In this paper we take advantage of the connectivity idea to locate the elbow and knee joint points from other known points derived in the previous steps. First we divide the joint set into two groups: hand-elbow-shoulder and feet-knee-hip. In Figure 5, based on the human kinematic constraints, if we make a straight line L from shoulder (the green point) to hand (the blue point), the elbow (the black point) has a limited movement on the red line $L_\perp$ which intersects with L at the midpoint of shoulder to hand. Therefore, by searching all possible points on $L_\perp$, the elbow point can be obtained by minimizing the equation below:

$$E_{total} = \arg\min\{E(hand, elbow) + E(elbow, shoulder)\}, \; E \neq 0 \quad (3)$$

where E is the connectivity energy function. Same operation is performed on the hip-knee-foot set so that the complete human body model can be constructed.

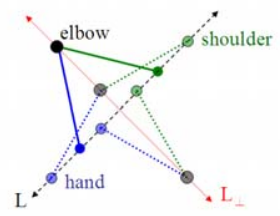

Figure 5: The kinematic constraint on the arm.

3. SIMULATION RESULTS

In our experiments, a set of human motions are tested for the feasibility of the algorithm. The videos are taken in an indoor environment with stable lighting condition and a fixed side-view camera. The ordered contour boundary points set are sampled every 8 points empirically after the VO segmentation based on background subtraction with 4-th order moment analysis. Four different types of behavior were conducted in our experiments: walking, sitting down, standing up and falling down. Figure 6 shows the modeling results on three postures: walking, sitting down and falling down. The notations below the figures are frame numbers. Table 1 tabulates the tracking and modeling accuracy on our test videos.

Table 1. The precision rate of the modeling results

behavior		hand	feet	false alarm	Precision (%)
Walk	hand	475	0	0	100%
	feet	0	850	0	100%
Sit down	hand	150	0	122	18.67%
	feet	0	459	0	100%
Stand up	hand	131	0	131	0%
	feet	0	423	0	100%
Fall down	hand	519	3	17	96.72%
	feet	0	921	0	100%

4. CONCLUSION

We propose a 2D human body modeling system which integrates the video object extraction, idea of connectivity energy for joint point refinement and a nearest neighbor tracking algorithm. The whole human body is modeled by 11 body parts, which are effectively identified and tracked. The system has shown promising capability on modeling a human under a single video camera shot. Our current system is highly silhouette dependent; therefore human is required to carry no object in order to avoid bad silhouette extraction. This will help some further research such as surveillance system or human behavior analysis.

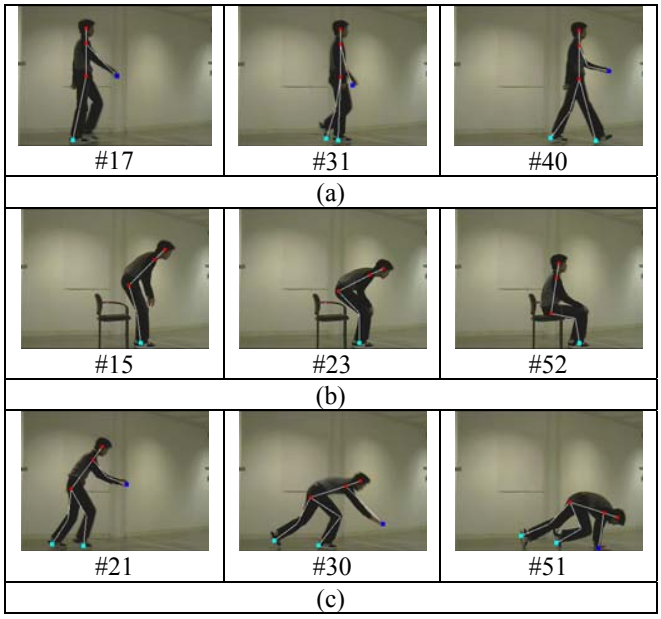

Figure 6: The modeling result of a (a) walking (b) sitting (c) falling sequence.

REFERENCES

[1] D. M. Gavrila. "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding: CVIU*, vol.73 (1), pp. 82–98, 1999.

[2] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. 9th Int. Conf. Computer Vision*, pages 734–741, 2003.

[3] G. Mori and J. Malik, "Estimating human body configurations using shape context matching." *In Proc. 7th European Conf. on Computer Vision*, Vol.3, pp. 666-680, 2002.

[4] S. X. Ju, M. J. Black, and Y. Yacoob, "Cardboard people: A parameterized model of articulated image motion." *2nd International Conference on Automatic Face and Gesture Recognition*, 1996

[5] X. Lan and D.Hutternlocher, "A unified spatio-temporal articulated model for tracking." *CVPR*, vol.1, pp 722-729, 2004.

[6] Fujiyoshi and A. J. Lipton. "Real-Time Human Motion Analysis by Image Skeletonization," *Proceedings of the Fourth IEEE Workshop on Applications of Computer Vision*, pp. 15-21, 1998.

[7] A. Neri et al., "Automatic Moving Object and Background Separation," *Signal Processing*, vol.66, pp219-232, 1998.

[8] D. Hoffman, W. Richards, "Parts of recognition," *Cognition* 18: 65-96, 1984.

[9] Siddiqi, K., & Kimia, B. B. "Parts of visual form: computational aspects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 239–251,1995

[10] L. Piegl, W. Tiller, "The NURBS Book, " Springer, ISBN 3-540-61545-8, 1997.

[11] Kalman, R. E. "A New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME - Journal of Basic Engineering* Vol. 82: pp. 35-45 , 1960