Model-based Human Motion Analysis in Monocular Video

W. W. Lok[1] and K. L. Chan[2]

Department of Computer Engineering and Information Technology,

City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

[1]50003587@student.cityu.edu.hk [2]itklchan@cityu.edu.hk

ABSTRACT

Tracking human motion in monocular video is a challenging problem in computer vision. It has found a wide range of applications such as visual surveillance, virtual reality, sports science, etc. This project aims to develop a model-based human motion analysis system that can track human movement in monocular image sequence with minimum constraint. No markers or sensors are attached to the subject. Given a clip of the video, the first step is to manually fit the 3D human model to the subject in the first frame of the video. Then background subtraction is used to extract the human silhouette. We propose the silhouette chamfer as the main matching feature. Chamfer distance measure is carried out on the extracted subject silhouette. The silhouette chamfer contains both the chamfer distance and region information. Finally, we use discrete Kalman filter to predict the pose of the subject in each image frame. The update step uses Broydent's method to optimize the predicted pose to fit the person's silhouette by using the cost function. We use the gait database SOTON to test our system. The image sequences contain human walking in both the indoor and outdoor environment. The motion tracking results demonstrate that our system has an encouraging performance.

1. INTRODUCTION

Video-based 3D human motion tracking is an important and challenging computer vision problem. The scope of this research area covers the detection, tracking, and perhaps interpretation of human movement in the image sequence. It has attracted many interests due to its wide range of potential applications. A visual surveillance system can detect people and monitor their activities. This can be used to provide security control in places such as car park. In sports science, the posture and gait analysis can help to train athletes and monitor their performance. Similar systems may also be used for medicine purpose. Realistic human body animation can also benefit from the knowledge of motion tracking. Typical applications include computer games, movie production, etc. Motion analysis is also useful in retrieval and automatic annotation of human activities in video database.

Many computer vision researchers have made great efforts in analyzing and recognizing human motion in image sequence. The video may be shot by one camera or by several cameras from different viewpoints simultaneously. The systems of Hogg [1] and Rohr [2] are specialized for a one degree of freedom (DOF) walking model. Edge and line features are extracted from images and matched to a cylindrical 3D body model. Wachter and Nagel [3] also use the model-based motion tracking approach. Body parts are modeled by right-elliptical cones. All the DOFs are determined by an iterated extended Kalman filter. Recently, Ning et al. [4] use a 12 DOFs 3D human model for motion tracking. Pose estimation relies on both boundary match and region match. A divide-and-conquer search strategy is adopted. Bregler and Malik [5] recover the 3D human motion information under the orthographic projection by marking the body segments in an initial frame. For the special complexity of human motion, the existing research methods lay much limitation on the human subject, such as a uniform and quiescent background, parallelism of human motion direction to the image plane, and skin-tight clothing of human [6].

In section 2, we will describe the 3D human model. The silhouette chamfer and model gradient feature extraction processes are described in section 3. The tracking process and the tracking fault correction are described in section 4. All results are shown and discussed in section 5 and finally a conclusion is drawn.

2. HUMAN MODEL

Our human body model consists of a kinematic skeleton of articulated joints controlled by angular joint parameters x. The 3D human body model is constructed of truncated cones. Each truncated cone represents one body part. The human body contains 12 rigid body parts, including torso, head, upper arms, forearms, thighs, calves and feet. The posture of a walker can be defined by a 12-dimensional vector:

$$x = (X, Y, \theta_{LS}, \theta_{LE}, \theta_{LH}, \theta_{LK}, \theta_{LA}, \theta_{RS}, \theta_{RE}, \theta_{RH}, \theta_{RK}, \theta_{RA})$$

where X and Y represent the coordinates of the global

position, θ is angular joint parameter (L=left, R=right, S=shoulder, E=elbow, H=hip, K=knee, A=ankle).

3. FEATURE EXTRACTION

3.1. Silhouette Chamfer

The distance transform (DT) has been applied in many image analysis tasks including shape description, feature detection, skeletonization, segmentation, and multi-scale morphological filtering. Generally, the chamfer distance with edge detection can be used suitably to measure the similarity between the model part and image data [7]. However, if we only consider the chamfer distance with edge as the only feature, it is insufficient in some circumstances. A typical example is shown in Figure 1. Although the model part is obviously wrongly fitted to the image data, the system still settles with this result due to the high matching score obtained by the pose optimization function. This phenomenon can be clearly illustrated by magnifying the image. A gap appears between the hand and the leg in the upper part of the image, while another gap appears between the legs in the lower part of the image. It is found that ambiguity easily appears when two body parts are close to each other.

To avoid such ambiguities, region information is considered in our approach. We propose the silhouette chamfer as the main matching feature. Chamfer distance measure is carried out on the extracted subject silhouette. The silhouette chamfer contains both the chamfer distance and region information. The chamfer algorithm searches for the best fit of edge points from two different images.

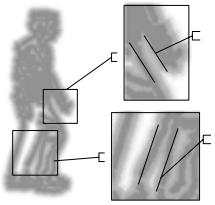


Figure 1 A typical ambiguity: a model part falls into the gap between two body parts in the image.

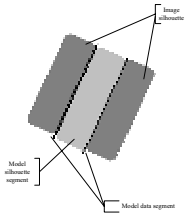


Figure 2 Model segment falls in between the body parts

In Figure 2, the model segment falls in between the body parts. Unlike using chamfer distance, this model segment is judged to be poorly fitted under the pose evaluation function when using the silhouette chamfer.

The silhouette image is extracted from the image sequence. The silhouette represents the presence of a feature. The feature templates are generated by the human model pose optimization process. They are grouped as the human model projected silhouette. The pixel values are the distances of the template features to the nearest features in the image. The lower the distance, the better is the result of matching the image to the template at this posture.

The distance from a predicted model point $p_i(x)$ to a given silhouette S_g can be determined by the chamfer measure between model prediction $p_i(x)$ and point s_i on the given silhouette S_g:

$$E_{si}\left(p_i(x), S_g\right) = \frac{1}{|p|}\sum \min_{s_i \in S_g}\left(p_i(x) - s_i\right) \quad (1)$$

If E_{si} is low, the human model is satisfactory matched. The gradual and long-range characteristics make the chamfer distance mapping a suitable evaluation measure to guide the search process. The overlapping area and chamfer distance stabilize the estimation and drive it towards the desired result. The cost term has the desired gradual property. It is good for recovering from global position tracking failure.

3.2. Model Gradient Information

The gradient of the model body part edge segment is assumed to be approximately equal to the corresponding image gradient. The model gradient at a position (x, y) depends on the distance d to the nearest model edge at state S.

A model edge segment has been projected onto the image plane. The model gradient magnitude can be computed as

$$h(x, y, S) = Ae^{-\frac{d}{2\sigma^2}} \quad (2)$$

where A is the contrast factor, which is set equal to the maximum expected gradient value. σ corresponds to the expected edge width.

The model gradient field h is approximately equal to the measured image gradient:

$$\nabla g(x, y) \approx h(x, y, S) \quad (3)$$

We use the gradient magnitude values in the measurement. The model gradient field h is approximately equal to the measured image gradient. The model gradient cost term is defined as

$$\mathrm{E_e} = \frac{1}{\mathrm{m}}\sum_{\mathrm{x,y}}\left(\nabla \mathrm{g(x,y)} - \mathrm{h(x,y,S)}\right)^2 \quad (4)$$

where m is the total number of the summation terms.

4. TRACKING PROCESS

Whether continuous or discrete, the optimization process is often carried out by the minimization of a cost function. Our cost function employs a combination of region and

edge information. We define the cost function as

$$\min_x f_c(I,x) = w_1 \frac{r_s}{r_s + r_g} E_S(I,x) + w_2 \frac{r_g}{r_s + r_g} E_g(I,x) + w_3 \Delta x \quad (5)$$

where I represents the image frame, E_s is the silhouette chamfer cost term, E_g is the model gradient cost term, r_s denotes the size of the silhouette image region, r_g denotes the size of the gradient image region, w_1, w_2 and w_3 are the weighting factors, and Δx are the changes of joint angles.

4.1. Similarity Measure

We present an operator M_s to measure the shape similarity between two binary images.

$$M_s = \frac{\left|\left(S_M \cap S_T{}^c\right) \cup \left(S_M{}^c \cap S_T\right)\right|}{\left|S_T + S_M\right|} \quad (6)$$

where $|\ |$ denotes the cardinal number of set. S_M is the set of pixels of the projected model silhouette and S_T is the set of pixels of the image object silhouette. Superscript c represents the inverse silhouette image.

It is known that M_s will be low when the silhouette of the model closely matches the object silhouette in the image.

4.2. Tracking Fault Correction

This measurement is used for correcting and smoothing the estimated joint trajectories. We compute the mean M_m and the standard deviation σ_s of the similarity measure for all frames.

The tracking fault appears when the similarity measure of a particular frame is larger than the threshold (Equation 7). The faulty joint angle is replaced by the linear interpolation of joint angles (Equation 8) of the previous and next frames.

$$M_s(t) > (M_m + \sigma_s) \quad (7)$$

$$\theta_{new}(t) = \left(\frac{\theta(t^+) - \theta(t^-)}{t^+ - t^-}\right) \times (t - t^-) + \theta(t^-) \quad (8)$$

$\theta(t)$ is the estimated joint angle at frame t. t^- represents the index of previous frame, and t^+ represents the index of next frame.

5. RESULT AND DISCUSSION

To verify the effectiveness of our approach, we have carried out a large number of experiments on video sequences with both indoor and outdoor scenes. We use the gait database SOTON [8] to test our human motion tracking system. The database is stored as a pre-cut digital video (DV) file format. Each record contains at least one complete gait cycle. A camera captures the image sequence with a stationary indoor or outdoor background at a rate of 25 frames per second and the resolution of 720 x 480 pixels.

The result (Figure 3) is shown from left to right representing the progress in time. The first row shows the raw image frames, while the second row shows the subject silhouette obtained by the background subtraction process. The third row shows the projected silhouette of the tracked model. The matching similarity measure is showed in the last row. The value close to 0 means a good similarity of matching. Conversely, the value close to 1 means a poor similarity of matching.

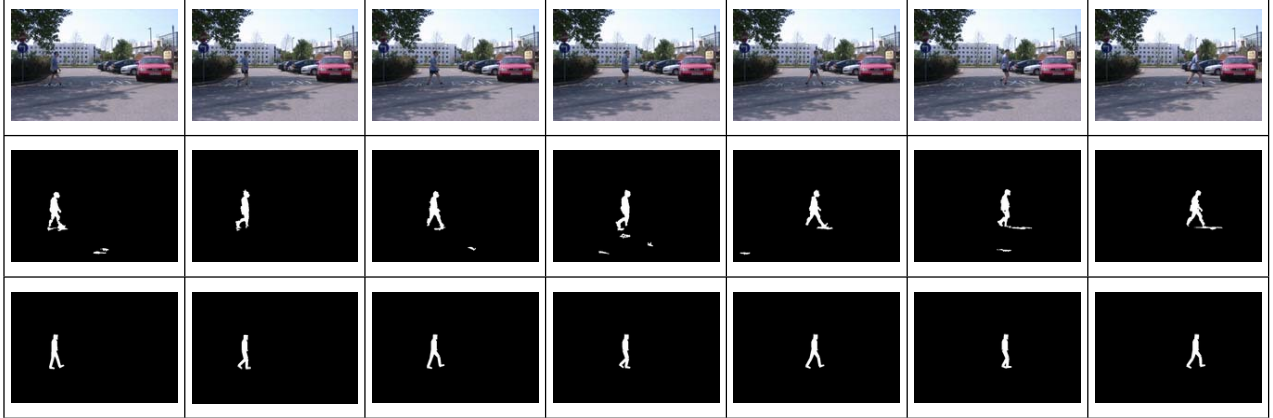
Frame 20	Frame 28	Frame 36	Frame 44	Frame 52	Frame 60	Frame 68
						
0.307862	0.226985	0.232636	0.352252	0.280102	0.373235	0.274257

Figure 3 Tracking result of outdoor scene 029e106s07R

Figure 4 shows the similarity measure. The mean of similarity measure is 0.2968 and the standard deviation is 0.0568. The tracking fault detection threshold is 0.3536. There are 11 faulty frames out of a total of 71. The fault rate is 15.5%. Large continuous tracking faults appear in frames 47-50. Figure 5 shows the tracked angles of the limbs. Each data point is an average of 3 frames. Tracking faults are marked in the figure. All trajectories move abnormally in frames 47-50. The left and right hip trajectories are intersected which means that two legs overlap.

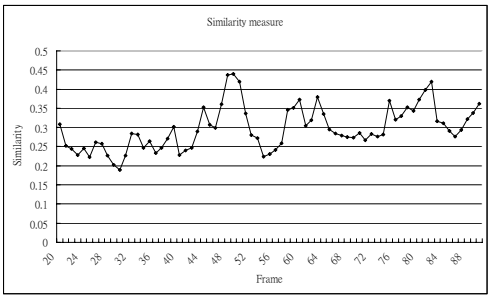


Figure 4 Similarity measure of outdoor scene 029e106s07R

The corrected limb trajectories are shown in Figure 6. The continuous tracking fault is improved from frames 47 to 50. The left hip continues to swing after the legs overlap. Similarly, right hip continues to lift. The limb trajectories are improved by applying the tracking fault correction.

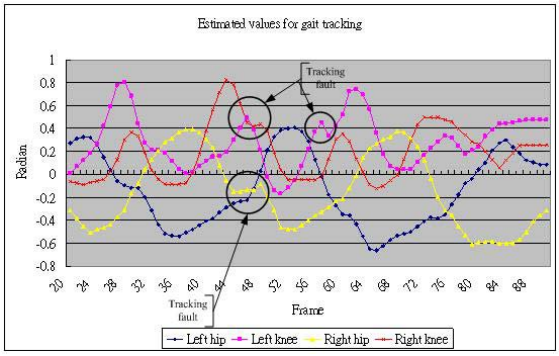


Figure 5 Tracked angles of the limbs

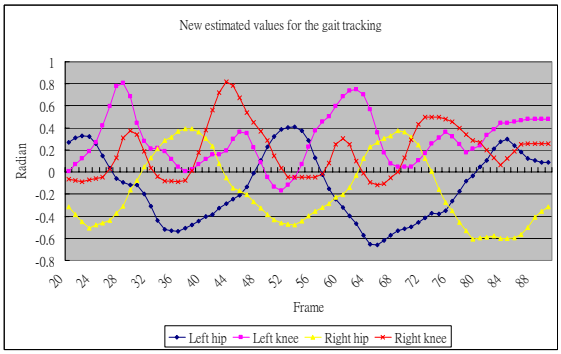


Figure 6 Tracked angles of the limbs with faulty values reassigned

The similarity measure provides important information for the identification of tracking faults. The tracking fault detection threshold is the sum of the mean and standard deviation of the similarity measure over the whole video. Based on this threshold, the faulty frames are identified. The joint angle of the faulty frame is recomputed and replaced by the linear interpolation of joint angles in neighboring frames. The adjusted limb trajectories are improved. Most of the continuous tracking faults appear when the legs overlap. By applying the tracking fault correction process, the tracking fault can be corrected. More results are available at www.it.cityu.edu.hk/~klchan/mphil.html.

6. CONCLUSION

We have demonstrated that our system is able to track human motion in a monocular image sequence. The gait database SOTON is used to test our system. The database contains the image sequences with a stationary indoor and outdoor background. The tracking fault correction amends the limbs trajectories and results in smooth detected motion.

7. REFERENCES

[1] D. Hogg, “Model-based vision: a program to see a walking person,” Image Vision Computing, 1(1), 1983, 5-20.

[2] K. Rohr, “Incremental recognition of pedestrians from image sequences,” Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, 1993, 8-13.

[3] S. Wachter & H. H. Nagel, “Tracking persons in monocular image sequence,” Computer Vision and Image Understanding, 74(3), 1999, 174-192.

[4] H. Ning, L. Wang, W. Hu & T. Tan, “Model-based tracking of human walking in monocular image sequences,” Proc. IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, 2002, 1, 537-540.

[5] C. Bregler & J. Malik, “Video motion capture,” Technical Report CSD-97-973, University of California, Berkeley, 1997.

[6] K. Aggarwal & Q. Cai, “Human motion analysis: a review,” Computer Vision and Image Understanding, 73(3), 1999, 428-440.

[7] D. M. Gavrila, “Pedestrian detection from a moving vehicle,” In D. Vernon, editor, Proceedings of 6th European Conference on Computer Vision, Dublin, Ireland, vol. 2, 37-49, June/July 2000.

[8] M. Nixon, J. Carter, J. Shutler & M. Grant, “Experimental plan for automatic gait recognition,” Technical report, University of Southampton, 2001.