

Model-based Tracking of Human Walking in Monocular Image Sequences

Huazhong Ning, Liang Wang, Weiming Hu and Tieniu Tan

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing, P. R. China, 100080

{hzning, lwang, wmhu, tnt}@nlpr.ia.ac.cn

Abstract We present a simple but effective method for recovering 3D human body walking parallel to the image plane from monocular video sequences based on 3D human body model, using robust image matching and specific search strategy. Monocular 3D body tracking is challenging, because of a large number of free parameters. These parameters are reduced to 12 in this paper under the assumption of walking parallel to the image plane. To effectively measure the matching error, a pose evaluation function combining both boundary and region information is provided. Simultaneously, we calculate the physical forces that are used as heuristic information in subsequent search. Experiments with both indoor and outdoor scenes demonstrate that our approach has an encouraging performance.

1 Introduction

Human tracking in video sequences has many potential applications such as visual surveillance, virtual reality, man-machine interaction, diagnostics, etc [1].

Our goal is to acquire 3D motion parameters including position, orientation and joint angles for gait recognition by tracking a walking person in an image sequence. Previous work on gait recognition mainly adopted 2D approaches [15, 16]. These methods often overlook the temporal features of joint angles which are much more essential in human walking. To achieve more robust and accurate results, we intend to use 3D motion parameters or combine both 3D and 2D features for gait recognition. As the important first step of our final goal, this paper considers 3D dynamic data acquisition.

Due to the complex nature of human body, tracking human in video sequences is a very difficult task [11] and involves a number of hard issues such as occlusion, self-occlusion, cluttered background, and high-dimensional motion parameters. To alleviate some of the difficulties, much previous work has used 3D human body models of various complexity to recover the position, orientation and joint angles from 2D image sequences (see surveys [1, 2, 8] for more information). In earlier research, stick figure model was frequently used [3]. The stick figure model is so simple that each human body part is represented by a stick and the sticks are connected by joints. More complex human models, such as cylinder [4, 9], truncated cone [5, 10] and super quadrics [6], were used in later work. Recently, Plankers and Fua [7] presented a hierarchical human model, which had four levels: skeleton, ellipsoid meatballs simulating tissues and fats, polygonal surface representing skin, and shaded rendering, to achieve more accurate results. As a general rule, the more complex the human body model, the more precise the tracking results. But on the other hand, complex human body model leads to extra computational cost. As a trade-off, it suffices for our purposes to adopt an articulated truncated cone human model with the head represented by a sphere.

Besides the human body model, image information used in pose evaluation function also varies. The most widely used image information is perhaps the boundary because it can be precisely localized and easily acquired [13, 4]. Another one is

the region which employs more information of the image and therefore achieves more robust results [14]. In this paper, we combine both boundary and region information in pose evaluation function to achieve both precision and robustness. To improve the speed of minimizing the pose evaluation function, we also design a specific search strategy that includes two stages: prediction and updating. In the prediction stage, the human global position is roughly estimated by finding the centroid of the human body, and the parameters of other body parts are predicted through computing the kinematical equation in the previous frame. In the updating stage physical forces similar to those in [5, 10] are used as heuristic information.

2 Human body model

Our human body model, similar to [5, 9, 10], consists of truncated cones (arms, legs, torso and neck) and a sphere (head). In detail, the human body is composed of 14 rigid body parts, including upper torso, lower torso, neck, two upper arms, two lower arms, two thighs, two legs, two feet and a head. Each body part is represented by a truncated cone except for the head which is represented by a sphere. Because hands are very complex and are relatively less important in human body tracking, we skip them to reduce the DOFs. Figure 1 gives some perspective projections of our human body model. This is a generic model and appropriate for average persons. But in the case of tracking a specific human, we must adjust its dimensions to individualize the model.

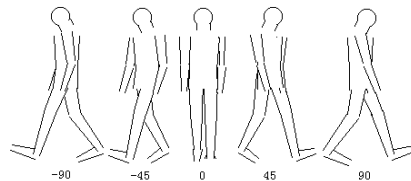


Figure 1. Human body model projected into the image plane from 5 viewing angles.

In general, a human body model has 45 DOFs: 3 DOFs for each body part (14×3) and 3 DOFs for its global position. To search robustly and quickly in the solution space with dimensions as high as 45 is almost impossible. However, in the case of gait recognition, human usually walks parallel to the image plane and the movements of the head, neck and lower torso, relative to the upper torso, are very slight. Therefore the solution space can be reduced with some constraints. In this paper, we assume that only the arms and legs have relative movements when the upper torso moves from left to right or from right to left. Furthermore, they move in the image plane, i.e. only joints such as shoulders, elbows, hips, knees and ankles are considered and each of them has only one DOF. Accordingly, DOFs of the human body model can be reduced to 12: 1 DOF for each joint mentioned above plus 2 DOFs for the global position. So the posture of a walker can be defined by a 12-dimensional vector $P = \{x, y, \theta_1, \theta_2, \dots, \theta_{10}\}$, where (x, y) is the global position, and θ_i is a joint angle. This vector describes the relative position of different body parts.

To locate the position of each model part, a local coordinate system is attached to it. The origin of the coordinate system at the upper torso indicates the global position of human body. The transformation of person coordinates into camera coordinate was detailed in [9].

3 Posture estimation

In our approach, the main task of model-based tracking is to relate the image data to the posture vector defined in Section 2. The general method to this problem is known as *analysis-by-synthesis*, and is used in *predict-match-update* fashion [1]. The philosophy, in detail, is to predict the most possible posture of the human body model in the next frame. Then the human model in predicted posture is projected into the image plane. We match the projected model to the edge image acquired by Sobel operator and measure the match error by a specific pose evaluation function. The remaining task is to find an optimal posture to minimize the match error with a specific search strategy to update the predicted posture according to previous match errors and previous frames. So the pose evaluation function and the search strategy play a vital role in the model based tracking. They are detailed in the following subsections.

3.1 Measuring boundary and region match

3.1.1 Boundary match

To make it easier, we match each body part independently at edge level, i.e. matching each body part of the projected model to the edge image acquired by Sobel operator. Figure 2 shows the matching procedure. For each pixel p_i in the boundary of the model part, we search the corresponding pixel in the edge image along the gradient direction at pixel p_i . In other words, the pixel, nearest to p_i and along that direction, is what we want. Assuming that q_i is the corresponding pixel and that F_i stands for the vector $\overrightarrow{p_i q_i}$, we can measure the match error of pixel p_i to q_i as the norm $\|F_i\|$. By averaging the match error of every pixel in boundary of the model part, the boundary match error is obtained

$$E_b = \frac{1}{N} \sum_{i=1}^N \|F_i\| \quad (1)$$

where N is the number of the pixels in that model part.

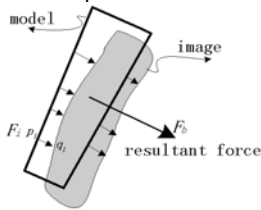


Figure 2. Measuring the boundary match and physical forces.

To provide the search procedure with heuristic information, we adopt the well-known idea of spring forces [5,10], which will be used in the next subsection but can be easily calculated here. With this idea, each F_i described above is viewed as a spring with its end points attached to p_i and q_i , and each spring gives a physical force in proportion to $\|F_i\|$, pulling p_i to q_i . Then the combination of all the physical forces, i.e. F_b in Figure 2, pulls the model part to the corresponding image part. The resultant physical force is given by

$$F_b = \frac{1}{N} \sum_{i=1}^N F_i \quad (2)$$

where N is the same as that in (1).

3.1.2 region match

Generally, the pose evaluation function defined in (1) can be used suitably to measure the similarity between the model part and image data, but it is insufficient under certain circumstances. A typical example is given in Figure 3(a), where a model part falls into the gap between two body parts in the edge image. Although it is obviously badly fitted, the model part may have a high matching score under the standard of the pose evaluation function (1). To avoid such ambiguities, region information is considered in our approach. Figure 3(b) illustrates the matching process. Here the region of a model body part, which is fitted into the image data, is divided into two parts: P_1 is the region overlapped with the image data and P_2 stands for the rest region. Then the match error with respect to the region information is defined by

$$E_r = |P_2| / (|P_1| + |P_2|) \quad (3)$$

where $|P_i|$ is the area, i.e. the number of pixels in the corresponding region.

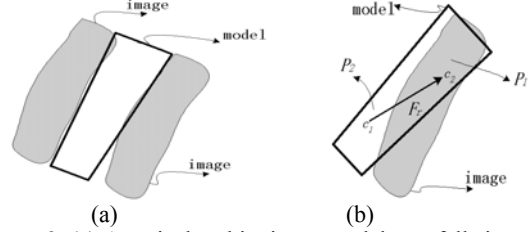


Figure 3. (a) A typical ambiguity: a model part falls into the gap between two body parts in the image. (b) Measuring region match and its physical force.

Similarly, another physical force is also defined. Supposing c_1 and c_2 are the centroids of the regions P_1 and P_2 respectively, we define the vector $\overrightarrow{c_1 c_2}$ as the physical force

$$F_r = \overrightarrow{c_1 c_2} \quad (4)$$

resulting from region matching. This physical force pulls the model part to overlap the corresponding image part as greatly as possible.

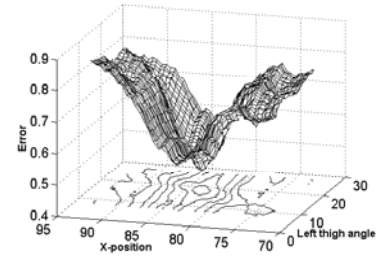


Figure 4. The curve of the pose evaluation function with the global position x and the joint angle of the left thigh changing smoothly and other parameters remaining constant. Also shown is the contour of the function.

To achieve a more proper measurement than (1) or (3) alone, both boundary and region match errors are combined, and so do the two physical forces. A factor α is used to adjust their weights.

$$E = \alpha \times E_b + (1 - \alpha) \times E_r \quad (5)$$

$$F = \alpha \times F_b + (1 - \alpha) \times F_r \quad (6)$$

How to determine the factor α is challenging. Enough experiments are needed to find an appropriate value of α . In our system, α is selected as 0.8 to improve the weight of

boundary match. Figure 4 shows the effectiveness of our pose evaluation function (5). Its curve is basically smooth and has no local minima at the neighborhood of the global minimum. Furthermore, according to the contour of the pose evaluation function and also other experiments, we can conclude that the human global position (x,y) is much more significant than other joint angles with respect to the match error. Also, this is the reason why the global position is first predicted without considering the change of other parameters, which will be detailed in our search strategy in the next section.

3.2 Search strategy

Locating an optimal posture in a high-dimensional body configuration space, e.g. 12 DOFs in our problem, is intrinsically difficult. To make it easier, a specific search strategy is designed to find a suboptimal posture. The search strategy includes two stages: prediction and updating. We describe them separately as follows.

In the prediction stage, the posture of the moving human in current frame is roughly estimated according to previous frames. To roughly estimate the global position, it is assumed that change of the centroid of the human body between consecutive frames is equal to the change of its global position

$$\begin{aligned} X_c - X_p &= C_c - C_p, \text{ or} \\ X_c &= C_c - C_p + X_p \end{aligned} \quad (7)$$

where the subscripts c and p indicate the current frame and the previous frame respectively, and X and C mean the global position and the center of gravity respectively. So predicting the global position of moving human in the current frame can be viewed as the problem of approximately calculating the center of gravity. To roughly predict the joint angles, we calculate their rotating velocities according to the previous frames and apply them to the kinematical equation

$$\theta_{ic} = \theta_{ip} + \dot{\theta}_{ip} \times \Delta t \quad (8)$$

where θ_i is the i th angle joint and the subscripts c and p indicate the current frame and previous frame respectively, and $\dot{\theta}_{ip}$ stands for the rotating velocity of that joint in the previous frame. Then θ_{ic} is the predicted value of that joint angle.

In the updating stage, we adopt the *divide and conquer* strategy to reduce the search space, i.e. we first update the global position of the human body in the current frame, and then update the joint angle of each body part one by one. The reason why we can update the global position and the joint angles separately is given in the previous subsection. To update the global position, we simply search at the neighborhood of the predicted global position to minimize the pose evaluation function. After the global position is determined, the positions of the shoulder and the hip are also fixed. So the movements of the upper arms and thighs can be considered as rotating around the fixed ends (shoulder or hip) in the image plane. So do the lower arms and legs after the upper arms and thighs fixed. We regard the physical force F described in (6) as the strength that acts on the centroid of the body part and pulls it to rotate. According to the rotating kinematical equation, we have

$$I\ddot{\theta} = rF \quad (9)$$

where I and $\ddot{\theta}$ are the moment of inertia and the joint angle acceleration respectively and r is the moment. But $I \propto L^2$ and $r \propto L$, where L is the length of the projected model part. So the joint angle acceleration $\ddot{\theta}$ is proportioned to F/L . After several steps of deduction using kinematical equations, the relationship between the change of the rotating angle and the physical force can be represented by

$$\Delta\theta = \beta F / L \quad (10)$$

where β is a scale factor determined by experiment. We update the rotating angle of each body part by adding the change in (10) to the predicted value or the last updated value.

To solve the problem of self-occlusion, we firstly locate the body parts near the camera (we know which body parts are near the camera if we know the walking direction) and save their projected image. Then when projecting the body parts far from the camera, we eliminate the occluded region according to the saved image. The method is particularly useful in the case of walking parallel to the image plane.

4 Experiments

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.

4.1 Data acquisition

For the indoor scene, we use the gait database SOTON from University of Southampton, UK, including seven subjects and four sequences for each subject. These sequences were captured by a camera with a stationary indoor background, at a rate of 25 frames per second, and the original resolution is 384×288 .

The sequences with outdoor scene with diffusing lightness was captured by a digital camera (Panasonic Nv-Dx100EN) fixed on the tripod at a rate of 25 frames per second. The original resolution of the images is 352×240 . These sequences form a portion of our NLPR gait database.

4.2 Results

We fit the human model to the image in the first frame manually and then the program tracks it automatically. Here we show two sequences as the tracking results (see Figure 5 and Figure 6). Due to the space constraint, only the human areas clipped from the original image sequences are shown. The most difficult part of the data, which verifies the effectiveness of our approach, is that the sequences include the configuration in which the two legs and thighs occlude each other severely (e.g. frame 11 in Figure 5 and frame 25 in Figure 6), causing most part of one leg or thigh is unseen. Other challenges include: images that have shadow under the feet; and the arm and the torso have the same color. It is worthwhile to mention that the arms far from the camera in both sequences were lost for the severe occlusion by the torso.

However, in our earlier experiments, region information is not considered. So when the human legs are in the special posture showed in Figure 7, the model part of right leg falls into the gap between the two legs and the tracking fails. By contrast, through the combination of the two match errors, the tracking results are much more robust (see frame 13 in Figure 5). This example clearly illustrates the effectiveness of our pose evaluation function.

Our purpose of the 3D model based tracking is to acquire 3D data of walking, such as joint angles and velocity, for gait recognition. We hope to find distinguishable characteristics of individual gait from those posture vectors of the tracking results. Figure 8 shows the temporal curves of hip and knee angles corresponding to the sequence in Figure 5 and Figure 6. Carefully inspection reveals that the periods in both sequences are about 34 frames. Further research on the analysis of 3D data of walking is an important part of our future work.

Although the proposed approach provides encouraging results, the work of this paper still has two considerable limitations. One is that our approach can hardly be applied to unconstrained movements because that the calculation and application of physical forces is based on the assumption that human walks parallel to the image plane. Another one is that we have to initialize the tracking process manually. To provide a general and really automatic approach to human motion capturing in any unconstrained environments, much work still

remains open.

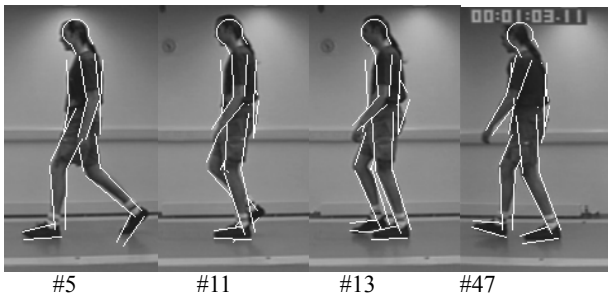


Figure 5. Tracking results of indoor walking.



Figure 6. Tracking results of outdoor walking.



Figure 7. Failure of tracking the right leg without considering region information.

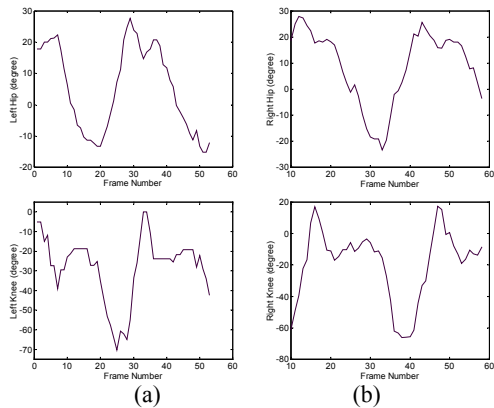


Figure 8. Temporal curve of joint angles. Top row: hip angles; bottom row: knee angles. (a) and (b) correspond to the sequences in Figure 5 and Figure 6 respectively.

5 Conclusion

We have presented our work on tracking of human walking parallel to the image plane in monocular image sequences based on 3D human body model composed of truncated cones and a sphere. An improved background subtraction technique was used to detect moving human from the background. Our main contribution is the provision of the pose evaluation function combining both boundary and region information, whose effectiveness was demonstrated by many experiments. Another contribution is that, to reduce the solution space, we presented a specific search strategy in *divide and conquer* fashion, i.e.

locating the global position and other body parts separately and using physical forces as heuristic information. The limitations of our approach were also discussed. The experiments on real sequences of both indoor and outdoor scenes show the effectiveness of our method.

Acknowledgements

The authors would like to thank Dr. M. Nixon and Ms. C. Yam from University of Southampton, U.K, for their help with the SOTON gait database. This work is supported by NSFC (Grant No. 69825105 and 60105002) and Institute of Automation (Grant No. IM01J02), Chinese Academy of Sciences.

References

- [1] D. Gavrilu, the Visual Analysis of Human Movement: a Survey, *Computer Vision and Image Understanding*, 73 (1), pp. 82-98, 1999.
- [2] T.B. Moeslund and E. Granum, A Survey of Computer Vision-Based Human Motion Capture, *Computer Vision and Image Understanding*, 81, pp. 231-268, 2001.
- [3] H.J. Lee and Z. Chen, Determination of 3D Human Body Posture from a Single View, *Comp. Vision, Graphics, Image Processing*, 30, pp. 148-168, 1985.
- [4] D. Hogg, Model-based Vision: A Program to See a Walking Person, *Image and Vision Computing*, 1(1), pp. 5-20, 1983.
- [5] Q. Delamarre and O. Faugeras, 3D Articulated Models and Multi-View Tracking with Physical Forces, *Computer Vision and Image Understanding*, 81, pp. 328-357, 2001.
- [6] C. Sminchisescu and B. Triggs, Covariance Scaled Sampling for Monocular 3D Body Tracking, in *Proc. of International Conference on Computer Vision and Pattern Recognition, CVPR'2001, Kauai, HI*.
- [7] R. Plankers and P. Fua, Articulated Soft Objects for Video-based Body Modeling, in *Proc. of 9th International Conference on Computer Vision (ICCV'2001), Vancouver, Canada*.
- [8] J. Aggarwal and Q. Cai Human Motion Analysis: a Review. *Computer Vision and Image Understanding*, 73 (3), pp. 428-440, 1999.
- [9] S. Wachter and H. H. Nagel, Tracking Persons in Monocular Image Sequences, *Computer Vision and Image Understanding*, 74(3), pp. 174-192, 1999.
- [10] Q. Delamarre and O. Faugeras, 3D Articulated Models and Multi-View Tracking with Silhouettes, in *Proc. of 7th International Conference on Computer Vision (ICCV'99), Kerkyra, Greece*.
- [11] J.C. Cheng, and J.M.F. Moura, Capture and Representation of Human Walking in Live Video Sequence, *IEEE Transactions on Multimedia*, 1(2), pp. 144-156, 1999.
- [12] T. Zhao, T.S. Wang and H.Y. Shum, Learning a Highly Structured Motion for 3D Human Tracking, in *Proc. of 5th Asian Conference on Computer Vision (ACCV'2002), Melbourne, Australia, 2002*.
- [13] D.M. Gavrilu and L.S. Davis, A 3-D Model-based Tracking of Humans in Action: a Multi-view Approach, in *Proc. of International Conference on Computer Vision and Pattern Recognition, San Francisco, CA*, pp. 73-80, 1996.
- [14] F. Lerasle, G. Rives, M. Dhome and A. Yassine. Human Body Tracking by Monocular Vision, in *Proc. of 4th European Conference on Computer Vision, Cambridge, England*, pp. 518-527, 1996.
- [15] J.D. Shutler, M.S. Nixon and C.J. Harris, Statistical Gait Recognition via Temporal Moments, in *Proc. of 4th IEEE Southwest Symposium on Image Analysis and Interpretation*, pp. 291-295, 2000.
- [16] J. Foster, M. Nixon, and A.P. Bennett, New Area Based Metrics for Gait Recognition, in *Proc. of 3rd International Conference on Audio- and Video-Based Biometric Person Authentication*, pp. 312-317, 2001.