3D Tracking of Human Locomotion: A Tracking as Recognition Approach *

Tao Zhao Ram Nevatia
University of Southern California
Institute for Robotics and Intelligent Systems
Los Angeles, CA 90089-0273
{taozhao|nevatia}@iris.usc.edu

Abstract

Estimating mode (walking/running/standing) and phases of human locomotion is important for video understanding. We present a new "tracking as recognition" approach. A hierarchical finite state machine constructed from 3D motion capture data serves as a prior motion model. Motion templates are used as the observation model. Robustness is achieved by making inferences in the prior motion model which resolves the short-term ambiguity of the observations that may cause a regular tracking formulation to fail. Experiments show very promising results on some difficult sequences.

1 Motivation and Introduction

Human motion tracking is important for a wide range of applications in motion recognition, human computer interaction and computer animation. In particular, tracking human locomotion is a key for human motion understanding and visual surveillance. High precision tracking can also be used for motion capture or gait analysis. In this paper, we present a *tracking as recognition* approach to recover the modes and phases of the human locomotion from monocular sequences; this effectively provides the coarse 3D joint angle trajectories of the human. A *mode* (e.g., walking, running, etc) is defined as a characteristic sequence of motion patterns or states which we call *phases*.

A large amount of effort has been put in human locomotion tracking (e.g., [3], [11]). A common formulation is to find the state at time t assuming the state at time $t-1$ is known. This formulation has a number of difficulties which prevent it from being robust in realistically noisy situations. Firstly, the state space is large for full-body motion (e.g., a human kinematics model has over 20 DOFs). Secondly, the track may drift when the observations (measurements) are ambiguous; such ambiguity is common when tracking a high dimension model using two dimensional image observataions; The Condensation [4] technique alleviates but does not completely solve this problem. Thirdly, the prior knowledge about the motion can only be used locally for prediction in tracking (e.g., [11], [8]). Finally, tracking requires knowledge of initial state; this is a difficult problem and a user often specifies it manually.

These difficulties call for stronger prior knowledge on the motion being studied and stronger temporal integration. We propose here a *tracking as recognition* approach where recognition of a locomotion mode assists in tracking. We employ a prior human locomotion model constructed from 3D motion capture data of human walking and running and represent it as a hierarchical finite state machine. The higher level state machine represents the modes of locomotion while the lower level state machine represents phases of each mode (walking and running only) with each phase corresponding to a characteristic pose. Tracking is performed as the following. In each frame, we measure the likelihood of each state using motion template. We buffer the observations for T frames and a best path in the motion model is computed by using the Viterbi algorithm. The best path gives the best mode and phase in each frame according to the observation of T frames.

In this approach, motion is inferred only in the prior motion model, which greatly constrains the solution space. Short-term ambiguities of measurements are resolved by considering all measurements globally. Initial state is estimated in the same way as all the other frames by inference. Besides, high level description (i.e., modes) is obtained at the same time. 3D body is estimated from a monocular sequence by using a 3D motion model.

In [10] pioneering work on motion recognition using medical motion data is presented. The joint angle values are represented as continuous curves and a search is done on the entire cycle to match the predicted edges with image edges. A Kalman filter is used to track the position and the phase information. The experiment is done only on human walking motion parallel to the image plane. In [7], a similar idea to treat tracking as an inference problem in an HMM is described; however, this approach is appearance-based and

[1]This research was supported, in part, by the Advanced Research and Development Agency of the U.S. Government under contract No. MDA-908-00-C-0036.

works well only for the viewpoints for which the system was trained. In [12], walking recognition was used to verify human hypotheses, however, the algorithm is computationally inefficient and constant walking velocity is assumed.

Due to the use of state-based motion model representation, the precision of the tracking in our method may be coarse for some applications such as motion capture. But this can still serve as the first step of a coarse-to-fine approach, in which the estimation can be refined locally from a good starting point. We believe that our tracking as recognition approach is a general formulation and can also be applied to other kinds of motion which can be well described with 3D limb trajectories and for which a state-based motion model can be built.

2 Tracking of Human Locomotion

This work is built on top of [12] where we presented a system to track the 3D positions of multiple humans in complex situations. Human segmentation and tracking is performed on the foreground obtained by statistical background subtraction. Known camera model and assumption of motion on a ground plane make tracking in 3D possible from a single camera. The 3D orientation of the human is inferred assuming he/she is facing in the direction of motion. In this work, we try to recover the mode of motion (e.g., walking/running/standing) and the detailed motion of the limbs using a tracking as recognition approach. We only track the motion of the legs, but the coarse motion of other parts is also obtained as we employ a full-body prior motion model.

2.1 The hierarchical locomotion model

Human locomotion has many modes, among which walking, running and standing are the three most often seen in daily life. A human can switch between these modes, therefore, the relationship of the modes is naturally represented as a finite state machine as shown in Fig.1.(a). The speed of the body is an important feature to distinguish among these three modes. The prior probability distribution of the speed given the modes $P(v|m), m \in \{walk, run, stand\}$ is set according to previous research in [1]. This finite state machine as well as the associated feature constitutes the first level of our hierarchical locomotion model.

A more detailed model is needed to track the more detailed motion of the limbs of walking and running. Walking and running are both periodic motions. We define a *cycle* to be the minimum repetitive unit, which equals to two *steps*. For each mode, several 3D motion capture sequences are gathered to compute an average cycle which starts from the phase when right leg crosses the left leg and moves forward. 3D motion capture data, consisting of a human kinematics model and a sequence of joint angle values, is a concise representation of motion of an articulated body and such data are now easily available. The average cycle trajectories are temporally quantized into a number of phases as the detailed representations of the modes.

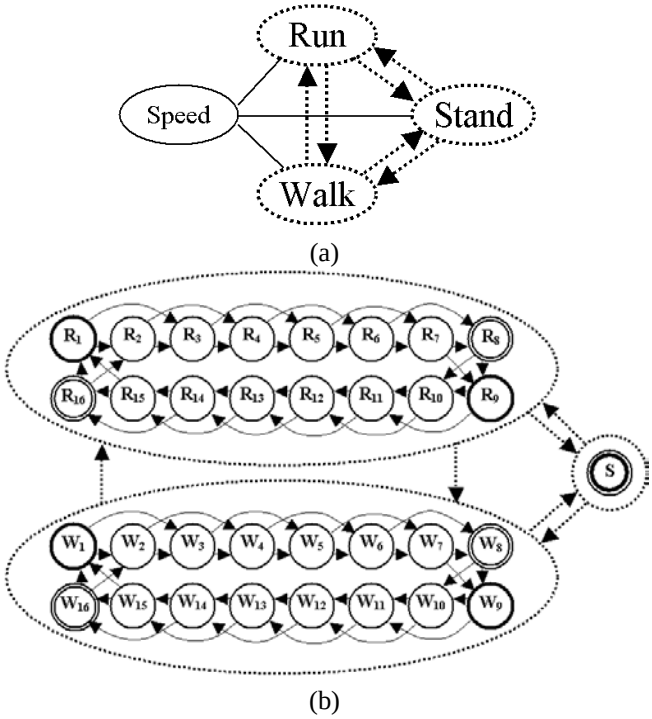


Figure 1: The human locomotion model. (a) the first hierarchy; (b) the second hierarchy (self-cycles omitted to save space). Notations: dotted ellipses and arrows-high level model; solid circles and arrows-low level model (specifically, dark circle-starting state; double circle-ending state); solid ellipses and solid lines-feature and its association.

Fig.1.(b) shows the finite state machine with walking and running modes expanded into their lower level state machines. 16 quantization levels are chosen considering the response property of the likelihood measurements (i.e., motion templates). Links back to states themselves (self-cycles) are not shown in the figure to save space. The bypass links (e.g., the link from R_1 to R_3) are added to handle fast motion. The start and end states in the low level state machine are defined so that the transition between modes can only happen at the start and end of each step. The transition probabilities within each mode are set according to average walking/running cycle duration, and the inter-mode transitions are set according to their observed frequencies. The initial state probabilities are set to be uniform on all possible states.

2.2 Motion Template

A distance measure of the state observables and image observations is needed for inference. Here we propose to use a *motion template* which is the template of model or image optical flow. It encodes both the shape and motion information. It is insensitive to the different texture of the object and has more discriminative power than static features such as a static template or edge. Therefore, it is suitable as a representation of instantaneous motion. Besides, its response field is relatively wide so that less quantization (e.g., 16 for each locomotion mode) is needed. It is different from the Motion Energy Image (MEI) and Motion History Image (MHI) introduced in [2] which are templates of accumulated motion over a period of time.

The motion template of the model can be easily computed from the 3D motion data given a camera model and 3D position and orientation of the human as in [12]. First the kinematics model in the motion data is rescaled according to the human height. The kinematics model is placed to the given position facing the given orientation. Forward kinematics is used to compute 3D positions of all joints and the camera model projects the 3D positions into 2D image positions.

With the 2D positions of the joints, the motion templates are computed as follows. For simplicity, we only show the procedure for one limb segment. We simplify the projected shape of the limb as an symmetrical (equicrural) trapezoid. Rotation along the limb direction is not considered since its effect is hardly visible in the image. Suppose we wish to compute the model flow at frame t with backward differencing. Let points A_t, A_{t-1} and B_t, B_{t-1} be the projected 2D positions of two ends of a limb segment in frame t and $t-1$ respectively. Their motions are:

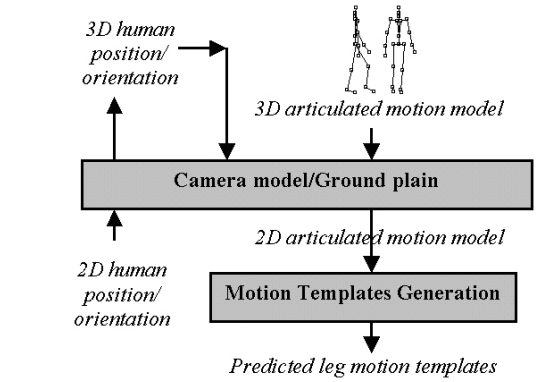
$$\Delta A = A_t - A_{t-1}$$
$$\Delta B = B_t - B_{t-1}$$

A point P_{t-1} in frame $t-1$ is transformed to its correspondence P_t in frame t in the following way (Fig.2.(b)). First it is translated by ΔA ($T_{translation}$), then it is rotated around A by θ ($T_{rotation}$), and then it is stretched along the direction of A_tB_t by a factor of s ($T_{stretch}$). θ is computed by cosine rule and $s = |A_tB_t|/|A_{t-1}B_{t-1}|$. Therefore, the motion of the point P is computed by Equ.1.

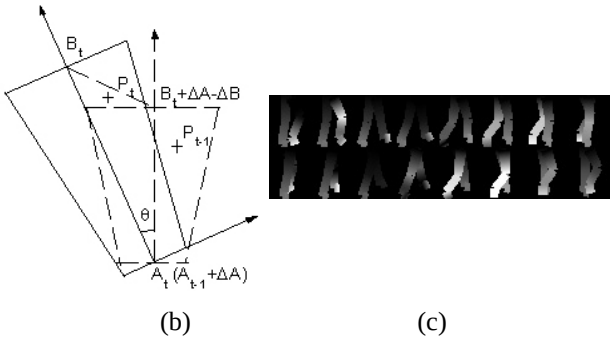
$$\Delta P = P_t - P_{t-1} = P_t - T^{-1}(P_t) \quad (1)$$
$$P_t = T(P_{t-1}) = T_{stretch}T_{rotation}T_{translation}(P_{t-1})$$

32 model motion templates (16 for walking and 16 for running) are generated this way in each frame.

Image motion or optical flow is computed only for the foreground objects from the incoming frames to avoid unnecessary computation. A block matching based optical flow algorithm is employed .



(a)



(b) (c)

Figure 2: Computing model motion template. (a) The diagram to compute motion template from model; (b) How the motion of each pixel is determined; (c) Result motion templates for walking at certain condition.

We define normalized template distance to be normalized sum of the vector differences of the two templates given by

$$D = \Sigma_{(x,y)\in\Omega}\frac{|\vec{v}_{xy} - \vec{m}_{xy}|}{|\vec{v}_{xy}||\vec{m}_{xy}|} \quad (2)$$

where Ω is the area within the bounding box of both legs, $\vec{v}_{xy}$ and $\vec{m}_{xy}$ are the vectors of image optical flow and the predicted motion template at (x, y) respectively. The distances are computed for the 32 model motion templates and the image optical flow aligning the feet of the model and the human in the image.

2.3 Inference in Motion Model

To perform tracking, we need to buffer the observations from frame 1 to frame T and to make inferences in the motion model to get an optimal path which maximizes the likelihood of the observations-both body speed and the motion template responses.

Denote λ as the motion model, $V = \{v_1, ...v_T\}$ as the speed, $O = \{o_1, ..., o_T\}$ as the motion template observations, and $Q = \{q_1, ..., q_T\}$ as the states in each frame. q_t is further decomposed into $[m_t, p_t]$ in which m_t is the mode and p_t ($p_t \in \{1, ..., 16\}$ for walking/running and

$p_t = 1$ for standing.) is the quantized phase in that state. We also number q_t from 0 to 32 where states $0, ..., 15$ are walking phases $W_1, ..., W_{15}$, states $16, ..., 31$ are running phases $R_1, ..., R_{15}$ and state 32 is the standing state. $A_{i,j} = P(q_t = j | q_{t-1} = i)$ is the state transition probability and we use $A_{0,i}$ to denote the initial state distribution for simplicity of notation. The normalized distance of motion template $[m_t, p_t]$ with the image optical flow is denoted as $D_t^{[m_t,p_t]}$. The optimal path is given by

$$Q^* = argmax P(O, V | Q, \lambda)$$

If we assume that the speed and the motion template responses are conditional independent, we have

$$\begin{aligned} P(O, V | Q, \lambda) &= \Pi_{t=1}^{T} A_{q_{t-1},q_t} P(o_t, v_t | q_t, \lambda) \\ &= \Pi_{t=1}^{T} A_{q_{t-1},q_t} P(ot | q_t, \lambda) P(v_t | q_t, \lambda) \\ &= \Pi_{t=1}^{T} A_{q_{t-1},q_t} P(o_t | [m_t, p_t], \lambda) P(v_t | m_t, \lambda) \end{aligned} \quad (3)$$

By assuming a Gaussian noise model ($N(0, \sigma)$, σ is the standard deviation of the Gaussian distribution; it is identical for all states) for the model motion templates, the likelihood of walking/running motion templates is given by

$$P(o_t | [m_t, p_t], \lambda) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{D_t^{[m_t,p_t]2}}{2\sigma^2}}$$

$P(o_t | m_t = stand, \lambda)$ is fixed to 1. The prior probability of the speed (second term in Equ.2) is given in Sec.2.1.

Computing the optimal path is similar to the problem of finding the best state sequence in an HMM. Therefore, we employ the Viterbi algorithm [9] based on dynamical programming for inference. A quantity $\delta_t(i)$ is defined as

$$\delta_t(i) = max_{q_1,..q_{t-1}} P(q_1, ..q_{t-1}, q_t = i, o_1, ..o_t, v_1, ..v_t | \lambda) \quad (4)$$

which is the maximum probability of states $q_1, ...q_{t-1}$ with q_t fixed to i. It can be computed from the first frame to the last frame by the induction rule:

$$\delta_t(j) = (max_{1 \le i \le N} \delta_{t-1}(i) A_{i,j}) P(o_t | q_t = j) P(v_t | q_t = j), \quad j = 1, ..., N$$

Another quantity $\psi_t(j)$ is defined as the previous state of a best path passing through state j, which is q_{t-1} in computing $\delta_t i$ in Equ.4. Its induction rule is:

$$\psi_t(j) = argmax_{1 \le i \le N} (\delta_{t-1}(i) A_{i,j}), j = 1, ...N$$

After $q*_T$ is computed, the best path Q^* is backtracked from frame $T-1$ to the first frame by:

$$q_t^* = \psi_{t+1}(q_{t+1}^*)$$

The resulting Q^* gives both the high level modes and the detailed phases which corresponds to body poses of each frame.

2.4 Optional post-processing

In many cases, the phase of walking or running changes approximately linearly with time. This linear relationship can be enforced in a post-processing stage. This can be used to deal with temporary missing/bad measurements due to occlusion, etc. For each segment of walking or running, the linear relationship is written as

$$\begin{aligned} p_t &= (p'_t)_{MOD16} \\ p'_t &= k * t + b \end{aligned}$$

The MOD operation destroys the direct linear relationship so that direct estimation of parameters k and b is not straightforward. The structure of the circular state machine restricts that the state can only move forward and can move at most one or two steps at a time. This property enables us to compute the p'_t before the MOD operation as

$$p'_t = p_t + n * 16$$

n is initialized to zero and is increased by one whenever p_t is less than p_{t-1} (e.g., $p_{t-1} = W_{16}$ and $p_t = W_1$). Then $\hat{k}$ and $\hat{b}$ are estimated from $\{p'_t, t\}$ pairs using LMS fitting. $\hat{p}_t = (\hat{k} * t + \hat{b})_{MOD16}$ is recomputed and the pose is interpolated by the poses corresponding to $\lfloor \hat{p}_t \rfloor$ and $\lceil \hat{p}_t \rceil$.

3 Experiment Results

3.1 Experiment Results

We have tested the proposed method on a variety of sequences, two of which are shown here. In both sequences, the 3D positions are tracked by the system in [12]. In our current implementation, all the frames are used for inference. In a real-time system, we can buffer a fixed number of frames and use a sliding window.

Sequence A is a 550-frame sequence in which a human first runs into the scene, switches to walking, stops at the corner for a while and walks back. Significant changes of orientation and size are visible in the sequence. Besides, the trajectory is not linear. Fig.3.(a) shows the estimated state of each frame in the optimal path and Fig.3.(b) shows the states after the option post-processing (note that the results in Fig.4 and Fig.5 show results before the post-processing step). The 3D human pose corresponding to the state is projected and overlain on the original video (some frames are shown in Fig.4). It shows that the estimated motion parameters are accurate and the transitions between modes

appear natural. The only inaccuracy occurs around frame 140 when the human's direction is almost aligned with the camera optical axis and the walking speed is low and the motion of the legs is not very salient. This inaccuracy is resolved in the post-processing step (Fig.3.(b)).

Sequence B is a 250-frame sequence from a soccer video. We only track one player who walks from left to right, stops and then runs from right to left. The field of view of the camera covers one third of the entire soccer field so players are very small (less than 20 pixels in height), the image quality is low and multiple shadows are present. These difficulties, together with the fast motion of running, make tracking very difficult. Our approach also gives good result on this data, and some of the frames are shown in Fig.5.(a). Cropped images are shown due to space limitations.

Since we can obtain 3D body poses by our approach, human model can be rendered from a new viewpoint. For the soccer game application, the user may choose a virtual viewpoint to watch the game as in [6]. Fig.5.(b) shows the 3D body pose recovered in two different viewpoints with a figure animation tool Poser with the figure body and clothing provided the tool.

3.2 Computation

The program is implemented in un-optimized C++ code and runs at about 5 Hz on a Pentium 4 1.5G Hz PC. The main computation effort is spent on reading image files, computing image optical flow and generating model motion templates. The time for inference part (i.e., the Viterbi algorithm) is negligible. We feel that the expensive steps can be optimized to run in real-time and that some of them can be performed in parallel. As a trade-off for computation, simpler but not as powerful observation model (e.g., shape, edges, etc) could also be used.

4 Conclusion and Future Work

We have presented our work on tracking human locomotion with a tracking as recognition approach. A hierarchical finite state machine constructed from 3D motion capture data serves as a prior motion model. Motion template is proposed as a powerful observation model. The tracking is done by making inference in the prior motion model. This resolves the short-term ambiguity of the observation that may cause failure in previous approaches. Experiments have shown very promising results even on very difficult sequences.

There are a number of interesting directions in which this work can be extended:

- Human's orientation cannot be inferred from velocity when he/she is standing. Other information such as shape should be used to infer the orientation.

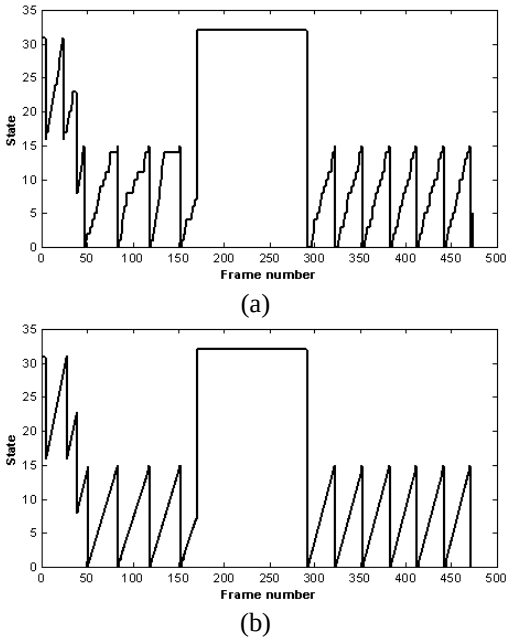


Figure 3: State output. 0-15 walking; 16-31-running; 32-standing. (a) before optional post-processing; (b) after optional post-processing.

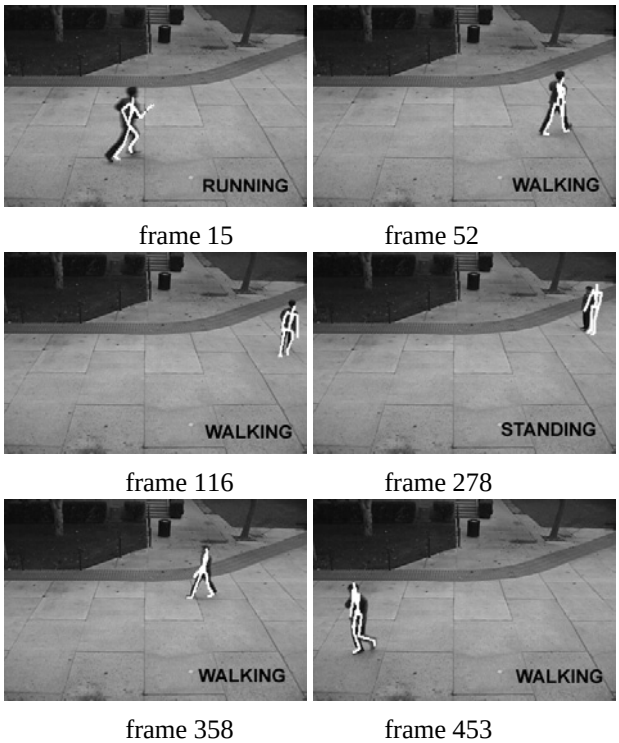


Figure 4: Result of sequence A. Key frames of the tracking results overlaid on images.

- More experiments can be carried out on other kinds of motion (e.g., dancing).
- In the motion where human's orientation is difficult to infer before hand, it is an interesting problem to estimate the orientation and motion phases at the same time.
- The motion we recover is coarse and needs to be refined for some purposes such as motion capture or gait analysis. Motion parameters and body parameters can be optimized locally to best fit the images.

References

[1] G. A. Bekey, Walking, *in The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, editor, MIT press, 1995.

[2] A. F. Bobick, J. W. Davis, The Recognition of Human Movement Using Temporal Templates, *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, 2001.

[3] C. Bregler, J. Malik, Tracking People with Twists and Exponential Maps, *Proc of IEEE Int'l Conf on Computer Vision Pattern Recognition*, Santa Barbara, CA, 1998.

[4] M. Isard and A. Blake, CONDENSATION - conditional density propagation for visual tracking, *Int'l Journal on Computer Vision*, 29, 1, 5-28, 1998.

[5] F. Lv, T. Zhao, R. Nevatia, Self-Calibration of a Camera from Video of a Walking Human, *Proc of Int'l Conf. on Pattern Recognition*, Quebec City, Canada, 2002.

[6] K. Matsui, M. Iwase, M. Agata, T. Tanaka and N. Ohnishi, Soccer Image Sequence Computed by a Virtual Camera, *Proc of IEEE Int'l Conf on Computer Vision Pattern Recognition*, Santa Barbara, CA, 1998.

[7] N. Krahnstover, M. Yeasin, R. Sharma, Towards a Unified Framework for Tracking and Analysis of Human Motion, *Proc of Int'l Workshop on Detection and Recognition of Events in Video*, Vancouver, Canada, 2001.

[8] V. Pavlovic, J. M. Rehg, T. Cham, A dynamic Bayesian network approach to tracking using learned switching dynamic models, *Proc of Int'l Workshop on Hybrid Systems: Computation and Control*, Pittsburgh, PA, 2000.

[9] L. R. Rabiner, A Tutorial on Hidden Markov Models and Slected Applications in Speech Recognition, *Proc of IEEE*, vol.77, no.2, 1989.

[10] K. Rohr, Towards Model-Based Recognition of Human Movements in Image Sequences, *CVGIP: Image Understanding*, vol.59, no.1, pp.94-115, 1994.

[11] H. Sidenbladh, M. J. Black, and D. J. Fleet, Stochastic tracking of 3D human figures using 2D image motion, *Proc of European Conference on Computer Vision*, Dublin, Ireland, 2000.

[12] T. Zhao, R. Nevatia, F. Lv, Segmentation and Tracking of Multiple Humans in Complex Situations, *Proc of IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, Kauai, HA, 2001.

[13] T. Zhao, T. Wang, H. Shum, Learning Highly Structured Motion Model for 3D Human Tracking, *Proc of Asian Conference on Computer Vision*, Melborne, Australia, 2002.

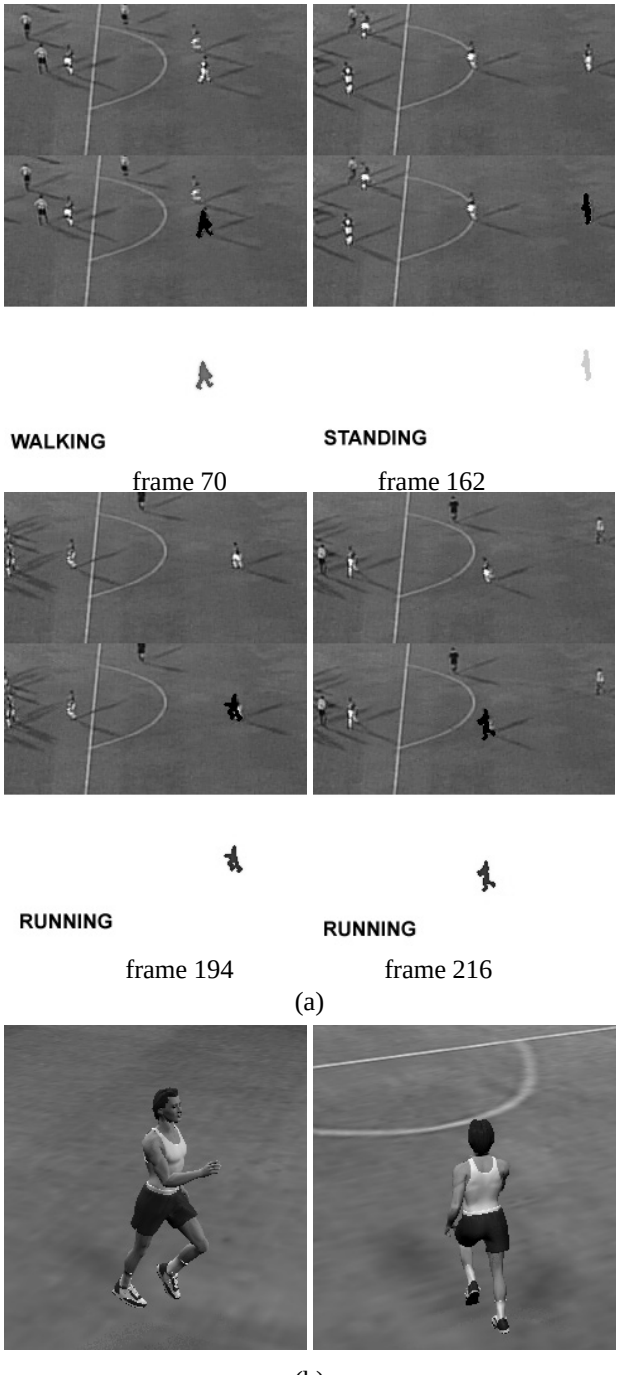


Figure 5: Result of sequence B. (a) Key frames of the tracking results overlaid on images. First row-original image; second row-model overlaid on original image; third row-model only. (b) Frame 194 as rendered from two new viewpoints using figure animation tool Poser.