TRACKING OF HUMAN BODY JOINTS USING ANTHROPOMETRY

A. Gritai and M. Shah

School of Electrical Engineering and Computer Science
University of Central Florida

ABSTRACT

We propose a novel approach for tracking of human joints based on anthropometric constraints. A human is modeled as a pictorial structure consisting of body landmarks (joints) and corresponding links between them. Anthropometric constraints relate the landmarks of two persons if they are in the same posture. Given a test video, where an actor performs the same action as in a model video, and joint locations in the model video, anthropometric constraints are used to determine the epipolar lines, where the potential joint locations are searched in the test video. The edge templates around joints and related links are used to locate joints in the test video. The performance of this method is demonstrated on several different human actions.

1. INTRODUCTION

Tracking of human joints is one of the important tasks in computer vision due to the vast area of applications. These applications include surveillance, human-computer interaction, action recognition, athlete performance analysis, etc. Joints tracking is a hard problem, since the appearance changes significantly due to non-rigid motion of humans, clothing, view point, lighting etc., therefore, appearance alone is not enough for successful tracking. We propose a novel approach for 2D joints tracking in a single uncalibrated camera using anthropometric constraints and known joint locations in a model video.

There has been a large amount of work related to this problem, and for a more detailed analysis we refer to surveys by Gavrila and Moeslund [2, 5]. The advanced methods are based on sophisticated tracking algorithms. The Kalman filter has been used previously for human motion tracking [8, 7], however, the use of the Kalman filter is limited by complex human dynamics. A strong alternative to the Kalman filter is the Condensation algorithm [4], employed by Ong in [6] and by Sidenbladh in [9]. In [1], Rehg modified the Condensation algorithm to overcome the problem of a large state space required for human motion tracking. However, even if a kinematic model is known, it is a non-trivial task to predict possible deviations from the model.

Since, humans perform actions with significant spatial and temporal variations that are hard to model, a tracker should take in account all aspects. Compared to some complex methods, our approach does not require specific knowledge in modeling human dynamics. Given a model of an action from any viewpoint, this paper proposes a novel approach to track joints in a single uncalibrated camera. Our motivation was the recent successful application of anthropometric constraints in the action recognition framework [3]. The anthropometric constraints establish the relation between semantically corresponding anatomical landmarks of different people, performing the same action, in a fashion, as epipolar geometry governs the relation between corresponding points from different views of the same scene. Because of the nature of anthropometric constraints, the epipolar lines, associated with landmarks, can slightly deviate from epipolar lines (due to the errors in positioning landmarks and linear relation between human bodies of different sizes). However, they still can reasonably approximate the landmark locations. Anthropometric constraints and known image positions of joints in a model video can be combined in as alternative approach to complex methods. As with previous methods, the proposed approach also has limitations, mainly due to view geometric constraints; however, these limitations can be solved without strong additional efforts. The performance of the proposed approach is demonstrated on several actions.

2. A HUMAN MODEL

We consider a window around the joint for modeling. This window provides us with the color and the edge information. The detection and tracking of joints can be improved by imposing constraints on their mutual geometric coherence, i.e. the optimal joint locations must preserve an appearance of the links (body parts) connecting joints. Image regions corresponding to links contain more essential information than windows around joints. Windows around joints and regions corresponding to links can be perfectly embedded in a pictorial structure. We refer to an entity performing an *action* as an *actor*. A *posture* is a stance that an actor has at a certain time instant, not to be confused with the actor's *pose*, which refers to position and orientation (in a rigid sense). The pose and posture of an actor in terms of a set of points in 3-space is represented in terms of a set of 4-vectors $Q = \{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^n\}$, where $\mathbf{X}^k = (X^k, Y^k, Z^k, \Lambda)^\top$ are ho-

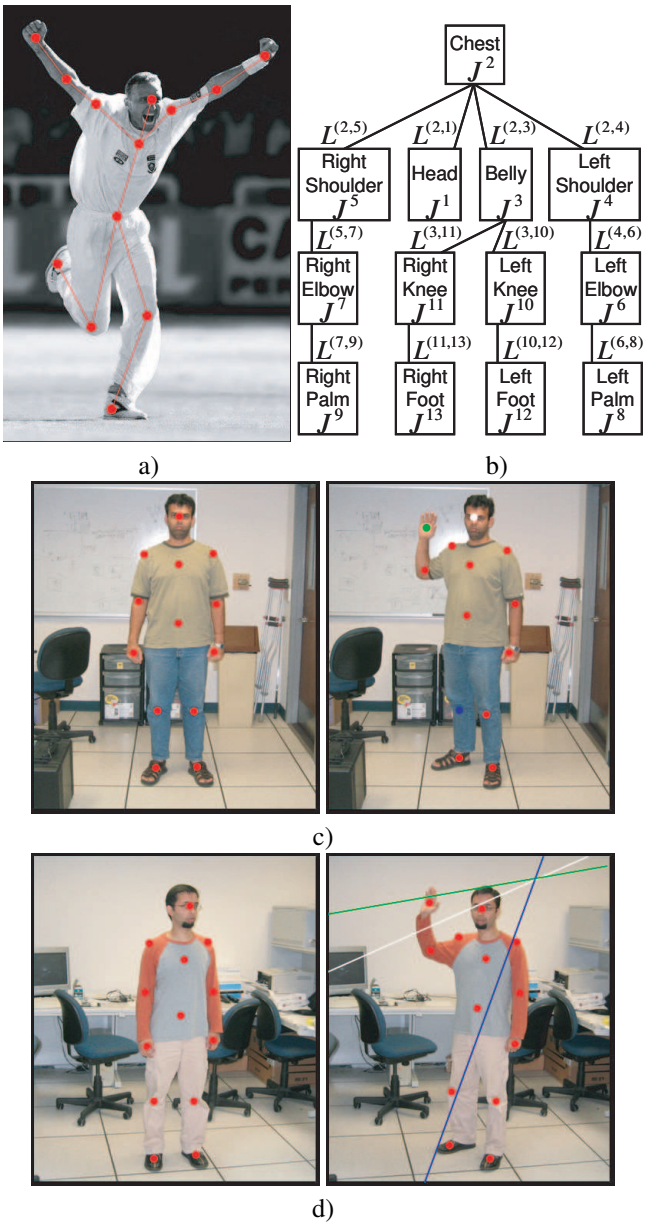

Fig. 1. a) Point-based representation. b) Pictorial structure showing joints and corresponding links. c-d) The fundamental matrix captures the relationship between joints of two different actors that are in the same posture and the variability in proportion as well as the change in viewpoint. c) An actor in two frames of the model video. d) Another actor in the corresponding frames of the test video. The joint correspondences in first frames of model and test video were used to compute the fundamental matrix. The image on right in d) shows epipolar lines in different colors corresponding to joints in the image on right in c). As it is clear that the joints in the test video lies on the corresponding epipolar lines.

mogenous coordinates of a joint k. Each point represents a spatial coordinate of a joint as shown in Fig.1 a), and points are connected by links. Thus, a human body is represented as a pictorial structure defined as follow

$$\mathbf{P} = (\mathbf{V}, \mathbf{S}),$$

where $\mathbf{V} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$ corresponds to joins, and $\mathbf{S} = \{\mathbf{L}^{(k,j)} \mid k \neq j; k, j \in \mathbf{V}\}$ corresponds to links. The imaged joint positions are represented by $q = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$, where $\mathbf{x}^k = (a^k, b^k, \lambda)^\top$. $\mathbf{X}^k$ and $\mathbf{x}^k$ are related by a 4×3 projection matrix $\mathbf{C}$, i.e. $\mathbf{x}^k = \mathbf{C}\mathbf{X}^k$. In [3], we proposed a conjecture, which states that there exists an invertible 4×4 non-singular matrix relating the anatomical landmarks (Q and W) of two actors, if they are in the same posture, *s.t.* $\mathbf{X}^k = \mathcal{M}\mathbf{Y}^k$. As a consequence of this conjecture, we have the following. First, if q and w describe the imaged positions of joints of two actors, a fundamental matrix $\mathcal{F}$ can be uniquely associated with $(\mathbf{x}^k, \mathbf{y}^k)$, i.e. $\mathbf{x}^{k\top}\mathcal{F}\mathbf{y}^k = 0$, if two actors are in the same posture, see Fig.1 c-d). Second, the fundamental matrix remains the same for all frames during the action as far as the actors perform the same action.

3. TRACKING

We assume a model video corresponding to different actions is available in the database, and joint locations in the model video are known. The problem then is given an unknown test video, we need to simultaneously decide, which action it is and determine frame to frame joint correspondences.

Suppose in a test and model video actors perform the same action. Known image location of the joint k in the frame i of the model video is denoted by $\mathbf{y}_i^k$, and unknown image location of the joint k in the frame j of the test video is denoted by $\mathbf{x}_j^k$. Assuming the joint locations in each frame i of the model video and an initial correspondence among joints, $w_1 = \{\mathbf{y}_1^1, \mathbf{y}_1^2, \ldots, \mathbf{y}_1^n\}$ and $q_1 = \{\mathbf{x}_1^1, \mathbf{x}_1^2, \ldots, \mathbf{x}_1^n\}$, between the first two frames of the model and test video are known, we propose an algorithm for the joints tracking in the test video. Since we know enough number of joint correspondences between two starting postures of both actors, the fundamental matrix, $\mathcal{F}$, can be recovered. Thus, k^{th} joint location in the frame i, $\mathbf{y}_i^k$, of the model video corresponds to the epipolar line, $\mathbf{l}_j^k$, passing through k^{th} joint location in some frame j of the test video. From fundamental matrix, $\mathcal{F}$, we can compute an epipolar line using $\mathbf{l}_j^k = \mathbf{y}_i^k\mathcal{F}$. Thus, knowing $\mathcal{F}$ and imaged joint locations in the model video, it is possible to predict the joint locations in each frame of the test video.

3.1. Locating joints in test video

Assume that joint correspondences between frames, f_i in the model and f_j in the test video, are known, therefore $\mathbf{y}_i^k\mathcal{F}\mathbf{x}_j^k = 0$. We can impose constraints on the search space of joint locations in frame f_{j+1} of the test video by using the known joint locations in frames f_{i+m} of the model video, where m is a length of the temporal window and $m = 0, \ldots, T$. For each joint, $\mathbf{x}_j^k$, the search space will be embedded between

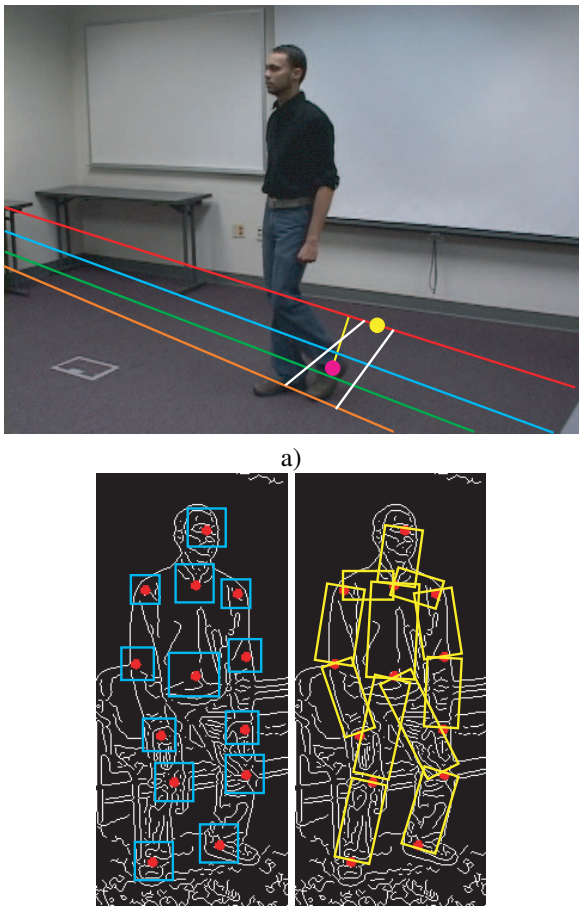

Fig. 2. a) The joint location in frame $\mathbf{f}_j$ is shown in yellow, and the correct joint location in frame $\mathbf{f}_{j+1}$ is shown in magenta. The epipolar lines shown in red, cyan, green and orange respectively correspond to the joint locations in frames $\mathbf{f}_i$, $\mathbf{f}_{i+1}$, $\mathbf{f}_{i+2}$ and $\mathbf{f}_{i+3}$ of the model video. The white lines constrain the deviation of the joint motion from the kinematic model. b) Left image shows the edge image and windows (edge maps) around joints. Right image shows regions corresponding to links connecting joints.

four lines. Two of them are epipolar lines corresponding to the joint locations in f_i and f_{i+m}, and the other two lines constrain the deviation of the joint motion from the kinematic model. In Fig.2 a), the location of the left foot in the previous frame of the test video is shown in yellow, and the correct location, which needs to be determine in the current frame, is shown in magenta. In this figure, the epipolar lines shown in red, cyan, green and orange, respectively correspond to the joint locations in frames f_i, f_{i+1}, f_{i+2} and f_{i+3} of the model video. As it is clear from the figure that none of the epipolar lines passes through the true joint location in the frame f_{j+1}. Therefore, we propose to search the true location of this joint in the space limited by epipolar lines and white lines, which constrain the deviation of the joint motion from the model.

The appearance model of each joint is represented by small (e.g. 16×16) window of the edge map centered around the joint location and its links in the first frame of the test video, see Fig.2 b). In order to find the match for the given joint in the current frame, we search for the location, which gives the minimum *Hausdorff* distance between the model template and the corresponding patches around the candidate location in the search space. Let $g(\mathbf{x}_1^k)$ and $g(\mathbf{L}_1^{(k,m)})$ respectively represents the edge maps around the joint k and its link to the joint m in the frame 1 of the test video. Then the *Hausdorff* distance between appearance model of the joint k in the frame 1 and its appearance in the frame j at some possible location, $\hat{\mathbf{x}}_j^k$, is denoted by $\mathbf{h}(g(\hat{\mathbf{x}}_j^k), g(\mathbf{x}_1^k))$. Similarly, the *Hausdorff* distance between appearance model of the link connecting joints k and m in the frame 1 and its appearance in the frame j is denoted by $\mathbf{H}(g(\hat{\mathbf{L}}_j^{(k,m)}), g(\mathbf{L}_1^{(k,m)}))$. Thus, the correct location of the joint k in the frame j of the test video is determined as

$$\mathbf{x}_j^k = \min_{\hat{\mathbf{x}}_j^k \in G_j^k} (h(g(\hat{\mathbf{x}}_j^k, \mathbf{x}_1^k) + \sum_{m \in N_k} H(g(\hat{\mathbf{L}}_j^{(k,m)}), g(\mathbf{L}_1^{(k,m)}))),$$

where G_j^k is a search space of the joint k in frame j, and N_k is a set of joints connected to the joint k.

The distance from the correct location of the joint k in the frame j of the test video to the epipolar line $\mathbf{l}_i^k$ of the corresponding joint location in the frame i of the model video is denoted as $d_{(j,i)}^k$. Similarly, the distance from the known location of the joint k in the frame i of the model video to the epipolar line $\mathbf{l}_j^k$ of the correct location of the joint k in the frame j of the test video is denoted as $D_{(i,j)}^k$. The correct correspondence between q_j in the frame j of the test video and w_i in the window of T-frames of the model video is determined as following

$$\min_{i \in T} \sum_{k=1}^{n} (d_{(j,i)}^k + D_{(i,j)}^k). \quad (1)$$

If there are several minima then the correct posture-state is the closest to the state i.

4. EXPERIMENTAL RESULTS

The proposed approach was tested on several actions including walking, sitting down, standing up, and standing up following by sitting down. Due to the limitation of space, the results of only two experiments, walking and sitting down, have been included here. In all experiments, the point correspondence was manually initialized between joints in the first two frames. The locations of all joints in the remaining frames were obtained automatically.

In the first experiment the model video was 125 frames long and contained one cycle of walking. The test video was 76 frames long. Fig.3 a) shows the tracking results in frames 3, 8, 17, 35, 47 and 75 of the test video. In the second experiment the model video was 79 frames long, and contained

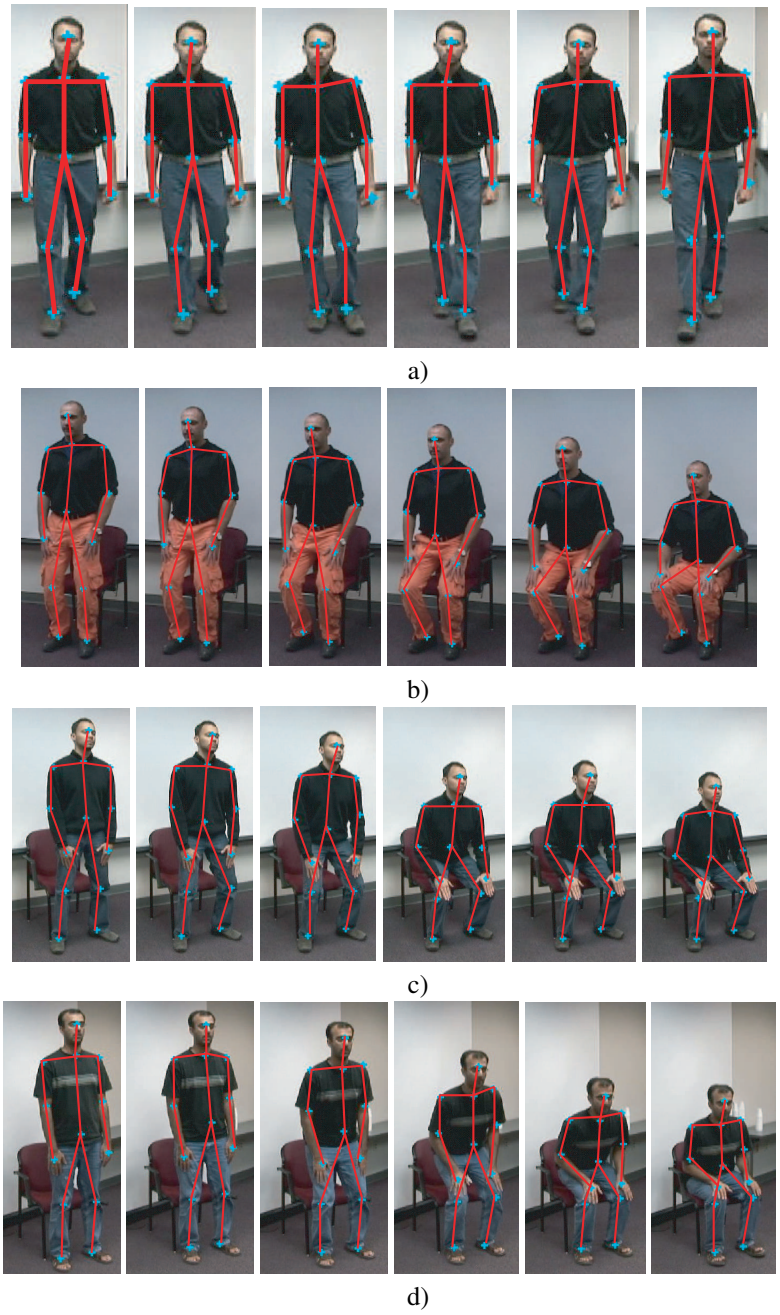

Fig. 3. a) The cyan marks show the joint location in the frames 3, 8, 17, 35, 47, and 75 of the test video. b) The cyan marks show the joint location in frames 3, 30, 56, 80, 113, and 172 of the first test video. c) The cyan marks show the joint location in frames 5, 19, 32, 43, 52, and 63 of the second test video. d) The cyan marks show the joint location in frames 4, 24, 36, 48, 60, and 72 of the third test video.

the example of "sitting down" action. Three test video were 180, 81 and 76 frames long. Fig.3 b) shows the joints tracking results in frames 3, 30, 56, 80, 113, and 172 of the first test video, Fig.3 c) shows the tracking results in frames 5, 19, 32, 43, 52, and 63 of the second test video, and Fig.3 d) shows the tracking results of joints in frames 4, 24, 36, 48, 60, and 72 of the third test video.

5. CONCLUSION

This paper proposed a novel approach for the tracking of human body joints. Compared to previous approaches, our method employed a much simpler model of human dynamics. The simplicity in modeling of human kinematics and good performance of the tracking should make the proposed method a promising alternative to the existing approaches. The performance of the tracking was demonstrated on several human actions.

6. REFERENCES

[1] T. Cham and J. Rehg, "Multiple hypothesis approach to figure tracking", *CVPR*, 1999.

[2] D. Gavrila, "The visual analysis of human movement: A survey", *CVIU*, 1999.

[3] A. Gritai, Y. Sheikh and M. Shah, "On the use of anthropometry in the invariant analysis of human actions",*ICPR*, 2004.

[4] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking",*IJCV*, 1998.

[5] T. Moeslund and E. Granum, "A survey of computer vision-based human motion capture", *CVIU*, 2001.

[6] E. Ong and S. Gong, "Tracking hybrid 2d-3d human models from multiple views", *International Workshop on Modeling People at ICCV*,1999.

[7] A. Pentland and B. Horowitz, "Recovery of nonrigid motion and structure", *PAMI*, 13(7):730742, 1991.

[8] N. Shimada, Y. Shirai, Y. Kuno and J. Miura, "Hand gesture estimation and model refinement using monocular camera - ambiguity limitation by inequality constraints", *CVPR*, 1996.

[9] H. Sidenbladh, F. De la Torre and M.J. Black, "A framework for modeling the appearance of 3D articulated figures", *International Conference on Automatic Face and Gesture Recognition*,2000.