

TRACKING OF HUMAN BODY JOINTS USING ANTHROPOMETRY

A. Gritai and M. Shah

School of Electrical Engineering and Computer Science
University of Central Florida

ABSTRACT

Most of the methods for human motion tracking are based on the modeling of human dynamics in action execution. In even small example space of human activities, the variation in action execution requires us to model a large number of uncertainties. This paper proposes a novel approach for motion tracking that avoids the tedious work of modeling human kinematics. This approach is based on the anthropometric and multi-view geometric constraints, successfully employed in the action recognition framework. 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. This paper proposes a novel approach for visual *2D* tracking of human body joints in a single uncalibrated camera using a known motion of joints 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 Gavrilu's and Moeslund's surveys,[2, 5]. Even if the kinematic model is known, it is a non-trivial task to predict a search space for human parts. The Kalman filter has been used previously for human motion tracking[8, 7], however, the use of the Kalman filter is limited by complex dynamics. A strong alternative to the Kalman filter is the Condensation algorithm[4] employed by Ong in[6] and by Sidenbladh in[9]. Rehg modified the Condensation algorithm in[1] to overcome the problem of a large state space required for human motion tracking.

People perform actions with significant spatial and temporal variations, and it is tedious work to model all variations. Compared to some previous methods, our approach does not require specific knowledge in the modeling of human dynamics. We propose an approach for *2D* human motion tracking in a monocular video using a known motion of joints in a model video. In a broad area of applications, the prior knowledge about human dynamics is a reasonable constraint. For example, a visual-based interface should expect an input as a known human action. Since there are variations in action execution, a tracker should take in account all aspects.

Similar problems arise in human action recognition. It is not unusual that the variety in human motion induces similar problems for human motion tracking and for human action recognition. Hence, we can apply some tools used for human action recognition to facilitate the motion tracking. Our motivation was the recent successful application of anthropometric constraints in the action recognition framework[3]. Anthropometric constraints allow us to use a person of average size as a reasonably good model for the human action. In addition to anthropometric constraints, the *2D* motion of human body joints, extracted from video, can be used as a kinematic model. It will be shown shortly how known *2D* motion of human joints, anthropometric and multi-view geometric constraints can be combined in an alternative approach to the Kalman filter and Condensation algorithm. As with previous methods, the proposed approach also has limitations, mainly due to multi-view geometric constraints; however, these limitations can be solved without strong additional efforts. We demonstrate the performance of the proposed approach on several actions.

2. A HUMAN MODEL

If we consider a human joint just as a point, there is not enough information to localize this point in the image. In our framework, a model of a joint is not just a point on the human body but a region around that point. The region around a body joint can provide us with color and edge information, and our experiments have shown that the edge information is more reliable. From this point forward, any reference to a joint means a centroid of the region around the body joint. The detection and tracking of joints can be improved by imposing constraints on their mutual geometric coherence. In other words, the optimal location of the joints must preserve an appearance of the human body parts connecting them. Image regions corresponding to human body parts, or links, contain even more essential information than regions around body joints. Regions around body joints and regions corresponding to their links connecting joints can be perfectly embedded in a pictorial structure.

To facilitate the further explanation of a human model, we follow the definitions presented in[3]. We refer to an entity performing an *action* as an *actor*. A *posture* is a stance

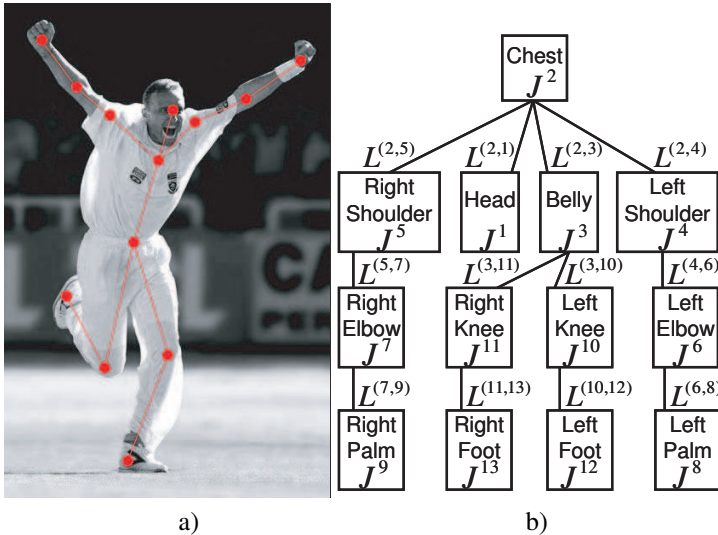


Fig. 1. (a) Point-based representation. (b) Pictorial Structure showing different body joints and their corresponding links.

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). An *action element* is the portion of an action that is performed in the interval between two frames. 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 $Q = \{\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^n\}$, where $\mathbf{X}^k = (X^k, Y^k, Z^k, \Lambda)^\top$ are homogenous coordinates of the joint k . Each point represents the spatial coordinate of an anatomical landmark on the human body as shown in Fig.1. Points are connected by links corresponded to human body parts. Thus, a human body is represented as a pictorial structure defined as follow

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

where $\mathbf{V} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$ corresponds to landmarks on a human body, and $\mathbf{S} = \{\mathbf{L}^{(k,j)} \mid k \neq j; j, k \in \mathbf{V}\}$ corresponds to links connecting landmarks. The imaged joint positions are represented by $q = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$, 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] an anthropometric constraint was used in recognizing human actions invariant to viewpoint, gender, race, sex, etc. The authors proposed a conjecture which states that there exists an invertible 4×4 non-singular matrix relating the landmark joint points (Q and W) of two actors, if they are in the same posture (stance that an actor has at certain time instant during an action) such that $\mathbf{X}^k = \mathcal{M}\mathbf{Y}^k$. As a consequence of this conjecture two results are immediate. First, if q and w describe the imaged positions of landmark points 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 the two actors are in the same posture (See Figure 2). Also, this fundamental matrix remain the same for all frames during the action as far as the actors

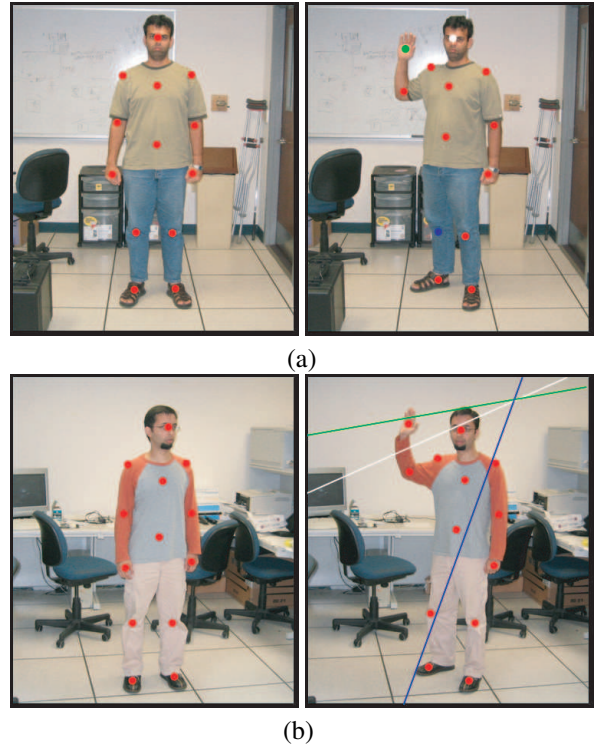


Fig. 2. The fundamental matrix captures the relationship between body joints of two different actors of different height, weight, etc. but in the same posture. The fundamental matrix captures the variability in proportion as well as the change in viewpoint. a) An actor in two frames of the model video. b) 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 (b) shows epipolar lines in different colors corresponding to joints in the image on right in (a). As it is clear that the joints in the test video lies on the corresponding epipolar lines.

are performing the same action. Fig.2 shows that the relationship between different postures can be captured by one fundamental matrix.

3. TRACKING

3.1. General approach

We apply anthropometric constraint to perform joint tracking. Assume we are given a test and a model video, in which 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 that the locations of all joints in the frame i , w_i , in the model video and an initial correspondence among joints, w_1 and q_1 , between the first two frames of the model and test video are known, we propose an algorithm for the tracking of body joints. Since we know enough number of

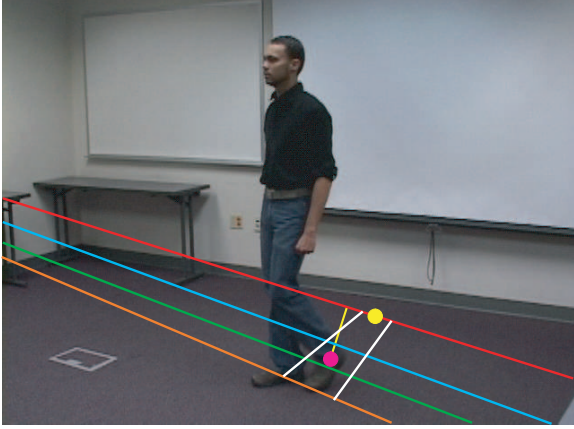


Fig. 3. The joint location in frame \mathbf{f}_j is shown in yellow, and the correct joint location in frame \mathbf{f}_{j+1} is shown in magenta. The epipolar lines shown in red, cyan, green and orange respectively correspond to the joint location in frames \mathbf{f}_i , \mathbf{f}_{i+1} , \mathbf{f}_{i+2} and \mathbf{f}_{i+3} in the model video. The white lines constrain the joint motion along epipolar lines.

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. Knowing \mathcal{F} , an epipolar line can be computed as $\mathbf{l}_j^k = \mathbf{y}_i^k \mathcal{F}$. Also, \mathcal{F} remains constant across the test video if two actors perform the same action. Thus, knowing \mathcal{F} and $2D$ joint location in each frame of the model video, it is possible to predict the joint locations in each frame of the test video.

3.2. Detection of joints

Assuming tracking of different body joints in different frames of the test video is known, anthropometric constraints can be used to recognize actions. However, as is well known, automatic tracking of body joint in a test video itself is a hard problem. In this paper, we show how the tracking of joints can be achieved using anthropometric constraints. We assume a model video corresponding to different actions are available in the database. The problem then is given a unknown test video, we need to simultaneously decide, which action it is and determine frame to frame joint correspondences. Assume that joint correspondences between frames, \mathbf{f}_i in the model and \mathbf{f}_j in the test video, are known, therefore $\mathbf{y}_i^k \mathcal{F} \mathbf{x}_j^k = 0$. Considering the fact that in a small window of T -frames the motion of joints is smooth, we can impose constraints on the search space of joint locations in frame \mathbf{f}_{j+1} in the test video by using the known joint locations in frames \mathbf{f}_{i+m} in the model video, $m = 0, \dots, T$. For each joint, \mathbf{x}_j^k , the search space will be embedded between four lines. Two of them are epipolar lines corresponding to the body joint locations in \mathbf{f}_i and \mathbf{f}_{i+m} , and the other two lines constrain the joint

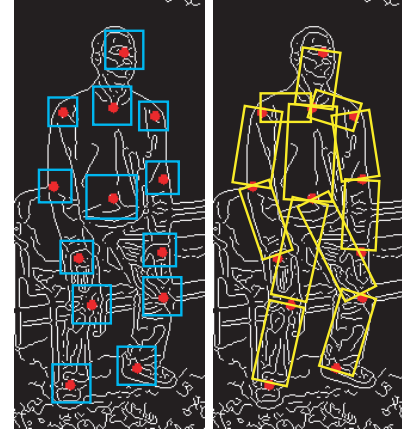


Fig. 4. Left image shows the edge image and bounding boxes (edge maps) around areas corresponding to joints. Right image shows edge maps around areas corresponding to links connecting joints.

motion along epipolar lines. In Fig.3 the location of the left foot in the previous frame in 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 location in frames \mathbf{f}_i , \mathbf{f}_{i+1} , \mathbf{f}_{i+2} and \mathbf{f}_{i+3} in 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 \mathbf{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 joint motion along epipolar lines.

The appearance model of each joint is represented by small (e.g. 16×16) patch of the edge map centered around the joint location and its links in the first frame of the test video, see Fig.4. 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 (patches around the joint and its corresponding links) 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 .

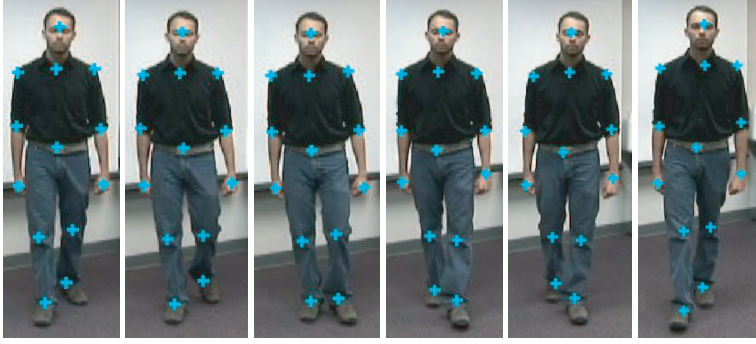


Fig. 5. The cyan marks show the joint location in the image.

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 a one cycle of walking. The test video was 76 frames long. Fig.5 shows the tracking results of joints 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 the test video was 180 frames long. Fig.6 shows the tracking results of joints in frames 3, 30, 56, 80, 113, and 172 of the test videos.

5. CONCLUSION

This paper proposed a novel approach for the tracking of human body joints. Compared to previous approaches, our method

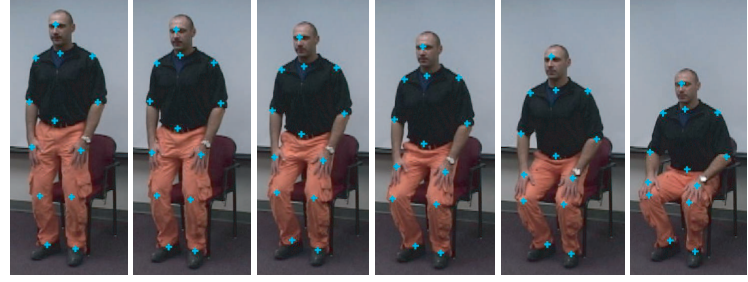


Fig. 6. The cyan marks show the joint location in the image.

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. Gavrilu, "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", *JCPR*, 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.