Geometric Constraints on 2D Action Models for Tracking Human Body

[†]Alexei Gritai, [‡]Arslan Basharat, and [‡]Mubarak Shah
[†]Cernium Corporation, Reston, VA, USA
[‡]Computer Vision Lab, School of EE & CS, University of Central Florida, Orlando, FL, USA
[†]agritai@cernium.com, [‡]{arslan, shah}@cs.ucf.edu

Abstract

We propose a 2D model-based approach for tracking human body parts during articulated motion. A human is modeled as a stick figure with thirteen landmarks, and an action is a sequence of these stick figures. Given the locations of these joints in a model video and only the first frame of a test video, the joint locations are automatically estimated throughout the test video using two geometric constraints. The first constraint is based on the invariance of the ratio of areas under an affine transformation, and provides initial estimates. The second one is based on the fundamental matrix, defined by the corresponding landmarks of the two actors, and refines the initial estimates. Using these estimated locations, the tracking algorithm determines the exact location of each joint in the test video. The novelty of our approach lies in the geometric formulation of human actions and the use of geometric constraints for body joints estimation. The approach is able to handle variations in anthropometry of individuals, viewpoints, execution rate, and style of action execution. Experimental results provide encouraging quantitative and qualitative performance analysis.

1 Introduction

The analysis of human motion and activities by a machine has attracted the attention of many researchers. Detection and tracking of different body parts (arms, legs, torso etc.) or landmark points (elbows, knees, shoulders etc.) provides important low level information for surveillance, human-computer interaction, action recognition, athlete performance analysis, etc. The success of the above mentioned applications strongly relies on the accuracy of body joint detection and tracking.

Since a general solution to body part tracking is considered difficult to find, approaches encapsulate constraints. These approaches are classified into *model-free* and *model-based* due to these constraints [1]. Model-free approaches do not rely on any prior knowledge about human pose or structure. Several bottom up detection approaches were proposed [3, 4, 5]. In these approaches body parts were detected using AdaBoost [3], 2D shapes [5], and local appearance models [4]. Model-free approaches may suffer from long computational time due to the need to prune wrong detections and human configuration [7]. Model-based approaches make use of the prior 2D and(or) 3D information about the structure and/or kinematics of human body. There were shape based approaches [2], motion based approaches [8, 10], and a combination of both types [9, 11]. Large databases of shapes and motion patterns [9, 6] increase robustness to viewpoint change. The analysis of a complex human motion requires features, which were extracted from the shape and the motion of human body [12, 6]. [13] proposes an approach motivated by the laws of physics, and presents a kinematic model. [14] presents a new approach, which is based on hierarchical learning of appearance features using an extensive training data set.

The proposed method explores underlying geometrical similarity between the model and test actions. The novelty of the method is the unique geometrical formulation of human action, and the combination of the two geometric constraints for estimation of the joint locations in the test video. One advantage of this work is the avoidance of error propagation from frame to frame in the estimation process, because each joint estimate is computed based on correspondence between first frame of the model and the test videos. Another advantage is that unlike most of the previous approaches, our approach separates spatial and temporal information, which allows for the recovery of spatial search space as soon as the human model is initialized in the first frame of the test video. The third advantage is a robustness to variations in anthropometry, execution rate, viewpoint and execution style. Another advantage is that the estimation phase helps detecting the cases of self occlusions. This is not computationally expensive and does not require extensive training.

2 Estimating Joint Locations

A human body is modeled by a stick figure, which connects 13 landmarks (head, neck, two shoulders, two elbows, two hands, belly button, two knees, and two feet). We assume that all the landmarks are available for the entire model video and only the first frame of the test video. Without loss of generality, the action execution rate in model and test video is assumed to be the same. This assumption will be relaxed in tracking phase.

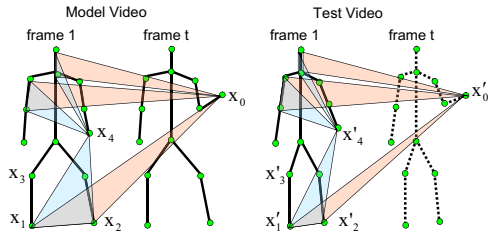

Figure 1. The unknown $\mathbf{x}'_0$ is estimated using known $(\mathbf{x}_i, \mathbf{x}'_i)$ and the ratio of areas, which are projections of triangles, e.g. $\triangle(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_4)$, $\triangle(\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2)$, $\triangle(\mathbf{X}'_1, \mathbf{X}'_2, \mathbf{X}'_4)$, and $\triangle(\mathbf{X}'_0, \mathbf{X}'_1, \mathbf{X}'_2)$.

Affine Constraint: A human action can be considered as a sequence of stick figures. We first show that the affine constraint can be derived for the human action in 3D, and then reduced to the 2D case in the image space. Fig. 1 shows stick figures from frame 1 and frame t of the model and test videos. The body points connected by solid lines are known, while those connected by broken lines (test video) are unknown, including $\mathbf{x}'_0$. Landmarks $\mathbf{x}_i$ represent 2D imaged locations of the 3D real-world landmarks $\mathbf{X}_i$. As shown in the figure $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$, and $\mathbf{X}_4$ can be considered coplanar in 3D with $\mathbf{X}_0$ off the plane. Connecting $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_4$ in 3D creates a volume $V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}$. Now consider volumes $V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}$ & $V_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}$ observed from model view and volumes $V_{\mathbf{X}'_0,\mathbf{X}'_1,\mathbf{X}'_2,\mathbf{X}'_4}$ & $V_{\mathbf{X}'_0,\mathbf{X}'_2,\mathbf{X}'_3,\mathbf{X}'_4}$ observed from test view. With the amount of out of plane motion being very small relative to the distance from the camera, an affine transformation relates the two volumes [16], i.e.

$$\frac{V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}}{V_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}} = \frac{V_{\mathbf{X}'_0,\mathbf{X}'_1,\mathbf{X}'_2,\mathbf{X}'_4}}{V_{\mathbf{X}'_0,\mathbf{X}'_2,\mathbf{X}'_3,\mathbf{X}'_4}}, \tag{1}$$

$$V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}=\frac{S_{\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}h}{3}, V_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}=\frac{S_{\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}h}{3},$$

$$V_{\mathbf{X}'_0,\mathbf{X}'_1,\mathbf{X}'_2,\mathbf{X}'_4}=\frac{S_{\mathbf{X}'_1,\mathbf{X}'_2,\mathbf{X}'_4}h'}{3}, V_{\mathbf{X}'_0,\mathbf{X}'_2,\mathbf{X}'_3,\mathbf{X}'_4}=\frac{S_{\mathbf{X}'_2,\mathbf{X}'_3,\mathbf{X}'_4}h'}{3},$$

where $S_{\mathbf{X}_i,\mathbf{X}_j,\mathbf{X}_k}$ is the area of base triangle and h is the height. Since $\triangle(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_4)$ and $\triangle(\mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4)$ lie on the same plane in 3D, the ratio of volumes in Eq.1 can be rewritten in terms of area, projected onto the image planes of both cameras

$$\frac{S_{\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}}{S_{\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}} = \frac{S_{\mathbf{X}'_1,\mathbf{X}'_2,\mathbf{X}'_4}}{S_{\mathbf{X}'_2,\mathbf{X}'_3,\mathbf{X}'_4}} = \frac{S_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_4}}{S_{\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4}} = \frac{S_{\mathbf{x}'_1,\mathbf{x}'_2,\mathbf{x}'_4}}{S_{\mathbf{x}'_2,\mathbf{x}'_3,\mathbf{x}'_4}}. \tag{2}$$

The ratios in Eq.1 and 2 can be also expressed in other terms

$$V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}=\frac{S_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2}h_1}{3}, V_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}=\frac{S_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3}h_2}{3},$$

$$V_{\mathbf{X}'_0,\mathbf{X}'_1,\mathbf{X}'_2,\mathbf{X}'_4}=\frac{S_{\mathbf{X}'_0,\mathbf{X}'_1,\mathbf{X}'_2}h'_1}{3}, V_{\mathbf{X}'_0,\mathbf{X}'_2,\mathbf{X}'_3,\mathbf{X}'_4}=\frac{S_{\mathbf{X}'_0,\mathbf{X}'_2,\mathbf{X}'_3}h'_2}{3},$$

where h_1, h_2, h'_1 and h'_2 are distances from $\mathbf{X}_4$ and $\mathbf{X}'_4$ to the planes of base triangles. Hence, the Eq.1 can be rewritten as

$$\frac{S_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2}h_1}{S_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3}h_2} = \frac{S_{\mathbf{X}'_0,\mathbf{X}'_1,\mathbf{X}'_2}h'_1}{S_{\mathbf{X}'_0,\mathbf{X}'_2,\mathbf{X}'_3}h'_2}.$$

Note that, for human articulated actions, the ratio of volumes in this equation can be approximated by area, projected onto image planes

$$\frac{S_{\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}}{S_{\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}} \approx \frac{S_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2}}{S_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3}} \approx \frac{S_{\mathbf{X}'_0,\mathbf{X}'_1,\mathbf{X}'_2}}{S_{\mathbf{X}'_0,\mathbf{X}'_2,\mathbf{X}'_3}}. \tag{3}$$

We estimate the location of the hand $\mathbf{x}'_0$ (see Fig. 1) using the constraints based on the invariance of ratio of triangular areas. One of these constraints can be derived from the area of triangles $\triangle(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4)$ and $\triangle(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2)$ from model video and another pair of $\triangle(\mathbf{x}'_1, \mathbf{x}'_2, \mathbf{x}'_4)$ and $\triangle(\mathbf{x}'_0, \mathbf{x}'_1, \mathbf{x}'_2)$ from the test video. Note that Fig. 1 shows the two frames side by side for illustration purposes only, in reality these triangles are projected on top of each other. The invariance of ratio of areas between the model and the test video is presented as

$$\frac{S_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2}}{S_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_4}} \approx \frac{S_{\mathbf{x}'_0,\mathbf{x}'_1,\mathbf{x}'_2}}{S_{\mathbf{x}'_1,\mathbf{x}'_2,\mathbf{x}'_4}} \Rightarrow S_{\mathbf{x}'_0,\mathbf{x}'_1,\mathbf{x}'_2} - \frac{S_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2}}{S_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_4}} S_{\mathbf{x}'_1,\mathbf{x}'_2,\mathbf{x}'_4} \approx 0.$$

This imposes one constraint for the solution of $\mathbf{x}'_0$ as a quadratic equation. Similarly, all other possible pairs of triangles, with $\mathbf{x}'_0$ as the common vertex, can be selected to apply more constraints on $\mathbf{x}'_0$. Since there are 13 landmarks, there are 66 possible triangle pairs. Thus, we have an over constrained system of quadratic equations of the form

$$S_{\mathbf{x}'_0,\mathbf{x}'_i,\mathbf{x}'_j} - \frac{S_{\mathbf{x}_0,\mathbf{x}_i,\mathbf{x}_j}}{S_{\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k}} S_{\mathbf{x}'_i,\mathbf{x}'_j,\mathbf{x}'_k} \approx 0,$$

where $k, i, j = 1, \ldots, 12$ and $k \neq i \neq j$. This system of equations is solved using nonlinear least squares. Note that recovering $\mathbf{x}'_0$ is independent of the other landmarks in the frame and does not rely on the computed locations in the previous frames, so the propagation of error is avoided. This estimation is only robust if an affine transformation relates the two viewpoints. In case of a perspective transformation, we need additional constraints to reduce the estimation error.

Epipolar Constraint: The estimation error induced by the affine constraint can be reduced by applying an epipolar constraint between the two actors. Given the correspondences among the joints in the first frame of the model and test videos, the fundamental matrix $\mathcal{F}$ is given by the relationship $\mathbf{x}_k \mathcal{F} \mathbf{x}'_k = 0$. A given joint location in the model video is mapped to a line in the test video through the fundamental matrix. The epipolar line limits the search space of the joint location in the test video. The point on the epipolar line closest to the initial estimate is chosen as the final estimate. The combination of the two constraints significantly reduces the estimation error for the pair of the wide baseline cameras, and captures the variations in viewpoints and the anthropometry of the individuals.

3 Joint Tracking

The estimated locations were used, along with the observed features, to track the 13 landmarks. In our method, foreground silhouettes are used as the observed features. We chose this simple feature to emphasize the importance of the previously computed joint estimates. This approach would be useful in challenging cases with non-discriminative local appearance (see Fig. 7). A human body is represented by a cardboard model, as shown in Fig. 2(h).

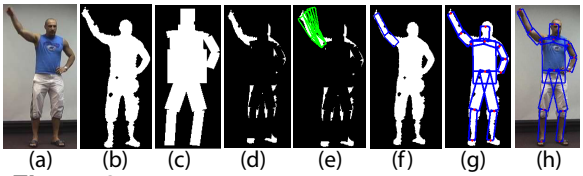

Figure 2. The main steps for detecting the right arm in the current frame. (a) The input frame of the test video, (b) extracted silhouette, (c) ICM for right arm, (d) subtraction of (c) from (b), (e) right arm templates from a subset of estimates are used to find position of the right arm, (f) selected best overlap, (g) the same procedure is repeated for the remaining body parts, and (h) final detection results.

In order to determine the exact locations of the body parts in the test video, each part of the cardboard model was fitted hierarchically to silhouette at each time instant. We agree with Navaratnam *et al.* in [7] that such hierarchical fitting is less complex than using the pictorial structure used in [2, 4]. The model is initialized in the first frame using the available locations of the 13 landmarks. For the following frames, the hierarchical detection of each segment in the test frame is performed in the following order: torso (belly, both shoulders), head, legs, and arms. The main steps involved in tracking the right arm are shown in Fig.2. Fig.2(a,b) show the current frame and extracted silhouette, respectively. The cardboard templates, positions of which were determined in the previous frame, constitute an intermediate cardboard model (ICM), in which the right arm is not included, see Fig.2(c). ICM is subtracted from the silhouette to isolate part of the silhouette corresponding to the right arm, Fig.2(d). The positions of the template corresponding to the right arm are drawn from a subset of estimates for each joint of the right arm over a temporal window, Fig.2(e). The best overlap between right arm template and isolated blob of the silhouette determines the best template position, Fig.2(f). The use of the temporal window addresses the variation in the rate of the action execution for each body part independently. The template location is further improved by a local search over template length and rotation angle. This operation accommodates for the spatial variation in the style of the action execution. Fig.2(g,h) show the final position of the cardboard model over the silhouette and current frame, respectively.

Occlusion Handling: A common problem in human body part tracking is self occlusion among body parts. Fig. 3(a) shows one example where both arms come in front of the torso and are not distinguished in the silhouette.

We use two measures for detecting start and end of self occlusion. The first measure is α_j^t, which represents the area of the foreground blob, corresponding to the j^{th} segment in the t^{th} frame. The second measure is β_j^t, which represents the proportion of the detected segment j that is occluded by the other segments of the cardboard model. The condition for occlusion is based on the normalized change over time τ,

$$\frac{\sum_{i=t-\tau}^{t-1} \alpha_j^{i+1} - \alpha_j^i}{\tau \alpha_j^{t-\tau}} < T, \frac{\sum_{i=t-\tau}^{t-1} \beta_j^{i+1} - \beta_j^i}{\tau \beta_j^{t-\tau}} > T,$$

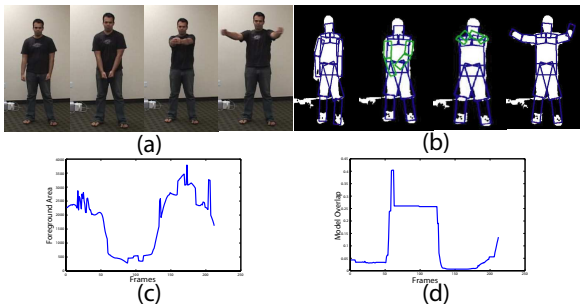

Figure 3. Occlusion handling: (a) Snapshots of the moving arms, (b) Cardboard models superimposed on silhouettes. Templates corresponding to occluded body parts are shown in green. Presence of occlusion is detected by the amount of change in α_j^t and β_j^t. In particular, (c) and (d) show α_j^t and β_j^t values corresponding to the left arm.

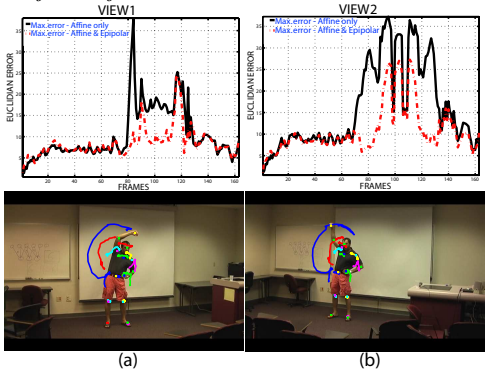

Figure 4. Error in joint estimation. The first row shows the maximum euclidian error in estimations from view 1 to 2 (a) and view 2 to 1 (b). The second row presents frames from each view with maximum error.

where T is the percentage threshold (we use 70% in our experiments) and τ is the size of the temporal window. Positive T value signifies entering occlusion, while the negative T value signifies exiting occlusion. The plots shown in Fig. 3(c,d) present the change in these parameters for the left arm. Once the start and the end of occlusion are determined, the difference in the rate of actions between model and test videos is calculated. Linear interpolation is used to estimate the joint locations during this occlusion interval. Fig. 3(d) shows the estimated location of the arms on the foreground silhouette during occlusion.

4 Experimental Results

Experiments were performed on several video sequences to analyze the performance of the proposed approach. These videos contained articulated motion, self occlusion, change in viewpoints, and a variety of actors. In the first experiment (shown in Fig. 4(a, b)) an action was captured by a pair of cameras with significant perspective variation. The length of the video was 163 frames. In both videos body joints were manually marked for the ground truth. Using the joint locations in view 1 and initialized landmarks in the first frame of the view 2, the locations of the joints in video 2 were computed and compared to the ground truth. This process was repeated in the reverse direction. Estimated trajectories were computed in two different ways: using only

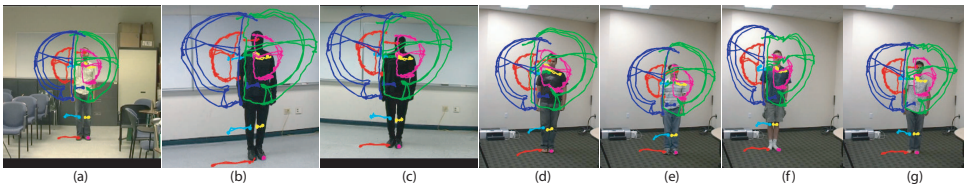

Figure 5. Joint estimation. (a) Trajectories corresponding to the model action. (b-g) show the trajectories of estimated joint locations of four different actors with significant changes in the anthropometric measurements.

Figure 6. The output of tracking on 285 frames sequence is shown here. The tracking output was observed to be accurate.

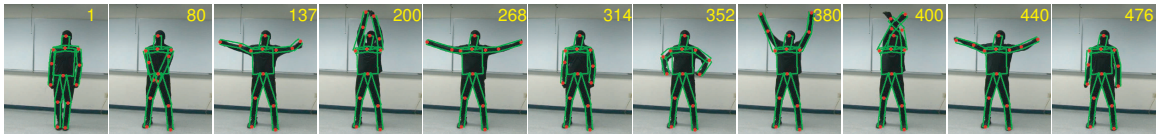

Figure 7. Output of the tracking phase on the video shown in Fig. 5(c). The test video is 476 frames long with articulated motion or arms. The arm undergoes full (frame 80) and partial occlusion (frame 400).

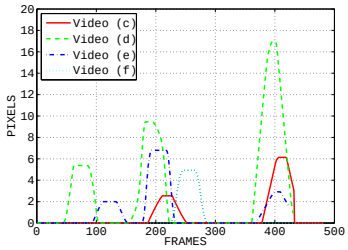

Figure 8. Plot shows mean of tracking error in the thirteen landmark locations. Each curve corresponds to a video shown in Fig. 5(c-f). Video (c) is also shown in Fig. 7 and the two peaks correspond to the instances with severe occlusion of arms.

affine constraint and affine with epipolar constraint. The first row of Fig. 4 shows the plot of maximum error in each frame. In this case the epipolar constraint reduces the error significantly. The second row of Fig. 4 shows the frames with the largest error. The results demonstrated here show the robustness to the viewpoint changes. Fig. 5 shows more estimation results on six videos with the predicted trajectories superimposed on a keyframe.

The results of the tracking phase are presented for two different actions. The results of the first action are shown in Fig. 6. This video contains 285 frames and the body parts were tracked correctly throughout all frames. The second action was more challenging as it contained larger variations in the viewpoint, anthropometry of individuals and the execution rate. In addition, there is large out of body plane motion and self occlusion. Fig. 5 shows the estimation results for this action and Fig.7 shows the tracking results for the video in Fig. 5(c). Fig. 8 presents a graph of the tracking errors from the four videos shown in Fig. 5(c-f). The peaks in error are at the point of self-occlusion.

5 Conclusion

We have proposed a novel 2D model-based approach for human body joint estimation and tracking. It can have variety of applications in the area of action and activity analysis. The formulation of the geometric constraints on the geometry of human action is novel. Compared to other approaches that use either linear or non-linear filters for human motion modeling, the proposed approach is easier to adapt to any model, more robust to viewpoint changes, and does not require extensive training. The experiments support the thesis that the proposed approach can handle significant variations in the anthropometry and execution rate. This research was funded by the US Government VACE program.

References

[1] T.B. Moeslund, A. Hilton, V. Krüger "A survey of advances in vision-based human motion capture and analysis" *CVIU* '06.
[2] P.F. Felzenszwalb, D.P. Huttenlocher,"Pictorial Structures for Object Recognition", *IJCV* '05.
[3] A. Micilotta, E. Ong, R. Bowden,"Detection and tracking of humans by probabilistic body part assembly", *BMVC* '05.
[4] D. Ramanan, D.A. Forsyth, A. Zisserman,"Strike A Pose: Tracking People by Findind Stylized Poses", *CVPR* '05.
[5] T.J. Roberts, S.J. McKena, I.W. Ricketts,"Human Pose Estimation using Learnt Probabilistic Region Similarities and Partial Configurations" *ECCV*, '04.
[6] S.X. Ju, M.J. Black, Y. Yacoob,"Cardboard People: A Parameterized Model of Articulated Image Motion", *FGR* '96.
[7] R. Navaratnam, A. T., P.H.S. Torr, R. Cipolla,"Hierarchical Part-Based Human Body Pose Estimation", *BMVC* '05.
[8] G.R. Bradski and J.W. Davis,"Motion segmentation and pose recognition with motion history gradients", *MVA* '02.
[9] A.F. Bobick and J.W. Davis,"The Recognition of Human Movement Using Temporal Templates", *PAMI* '01.
[10] H. Sidenbladh,"Detecting human motion with support vector machines", *ICPR* '04.
[11] B. Wu, R. Nevatia,"Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detection", *ICCV* '05.
[12] A.A. Efros and A.C. Berg and G. Mori and J. Malik,"Recognizing action at a distance", *ICCV* '03.
[13] M.A. Brubaker *et al.*,"Physics-Based Person Tracking Using Simplified Lower-Body Dynamics", *CVPR* '07.
[14] A. Kanaujia *et al.*,"Semisupervised Hierarchical Models for 3D Human Pose Reconstruction", *CVPR* '07.
[15] A.O. Balan *et al.* "Detailed Human Shape and Pose from Images", *CVPR* '07.
[16] R.I. Hartley, A. Zisserman,"Multiple View Geometry in Computer Vision", 2000.