View-invariant Recognition of Body Pose from Space-Time Templates

Yuping Shen and Hassan Foroosh
Computational Imaging Lab., University of Central Florida, Orlando, FL 32816
http://cil.cs.ucf.edu/

Abstract

We propose a new template-based approach for view-invariant recognition of body poses, based on geometric constraints derived from the motion of body point triplets. In addition to spatial information our templates encode temporal information of body pose transitions. *Unlike existing methods that study a body pose as a whole, we decompose it into a number of body point triplets, and compare their motions to our templates. Using the fact that the homography induced by the motion of a triplet of body points in two identical body pose transitions reduces to the special case of a homology, we exploit the equality of two of its eigenvalues to impose constraints on the similarity of the* pose transitions *between two subjects, observed by different perspective cameras and from different viewpoints. Extensive experimental results show that our method can accurately identify human poses from video sequences when they are observed from totally different viewpoints with different camera parameters.*

1. Introduction

Existing body pose recognition methods can be divided into two main categories: 3D approaches and 2D approaches. The 3D approaches try to recover the 3D body pose from 2D images and compare the body poses in the 3D Euclidean space. These approaches can be further divided into two groups: *model based approaches*, *e.g*. [5, 9, 13, 4], and *learning based approaches*, *e.g*. [1, 14, 11, 12]. A key problem in model-based approach is to define a good likelihood function in terms of edges [5, 15, 18, 17, 10], silhouettes [5, 9], intensities [16], or body joints [18]. Another important issue is how to reduce the cost search in the high-dimensional model parameter space. The learning based approaches avoid the high dimensional search in model parameter space by constraining the possible human poses to a small subset. To account for camera viewpoint variations, the training data usually consist of several 2D observations of plausible 3D poses.

Temporal information is also taken into account in some existing methods [8, 6] In this category, there are also a few *2D approaches* that do not require a 3D parametric body model or the prior knowledge of 3D configurations of body poses [3, 2]. However, these work do not take into consideration the variations in camera viewpoint and intrinsic parameters, thus limited in practice under wide camera variability.

1.1. Overview of Our Method

We represent the body pose by a space-time template: Spatial information is represented by a set of 2D imaged body points, and the temporal information by the transition between two poses. Unlike temporal templates proposed in [6] or snippets in [8] that require to be generated from multiple viewpoints, our template is generated from only a single arbitrary viewpoint, and the recognition is achieved using geometric constraints imposed by the motions of body point triplets, which we show to be invariant to camera calibration matrix and changes in viewpoints. View-invariance in most learning-based approaches is heavily dependent on the number of viewpoints used in the training data. Our solution eliminates these problems since only one viewpoint is required and the constructed likelihood function is independent of camera viewpoints and its intrinsic parameters.

2. Body Pose from Space-Time Templates

Our goal is to recognize a set of predefined poses $\mathcal{P}$ in a collection or sequence of 2D body poses $\{I_{1\dots n}\}$. The sequence can be from an uncalibrated camera and an unknown viewpoint. To this end, we maintain a set of space-time templates for a selected set of poses to be recognized. For each body pose, we require only one template from any arbitrary viewing angle. A body pose is recognized by matching the input unknown pose against all available templates and choosing the one with highest score, provided that the likelihood is above some threshold τ (see section 2.2). Our space-time templates require only the 2D image coordinates of body joints, and can be either extracted from a real video sequence or synthesized from motion capture data. A template is composed of an ordered pair of 2D poses: a key pose which represents the specific pose of interest and a succeeding one, which captures the transition shortly after the key pose. The succeeding pose can be selected arbitrarily, as long as it is sufficiently distinct from the key pose. Fig. 1 shows an example of a space-time template extracted from a tennis-serve video sequence.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

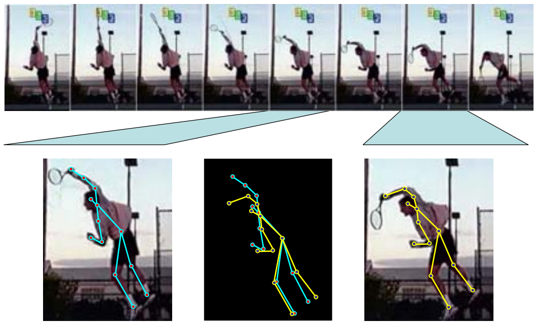

(a) (b) (c)

Figure 1. An example of a space-time template composed of two poses: (a) the key pose and (c) the succeeding pose. The two poses are overlapped in (b) to show their differences.

2.1. View-invariant Template Matching

Let I_i and I_j, $j > i$ be two frames with labeled body points in a video sequence, and T_1^k, T_2^k be the key pose and the subsequent pose of the template T^k, respectively. The key feature of our solution is the design of a matching score function $\mathcal{S}(I_{i,j}, T_{1,2}^k)$, that is independent of viewpoint and camera intrinsic parameters. Therefore, $\mathcal{S}(I_{i,j}, T_{1,2}^k)$ attains its maximum value when I_1 and I_2 correspond to T_1^k and T_2^k, regardless of camera parameters and varying viewpoints. The next section describes the steps to achieve this.

2.1.1 Template matching based on body point triplets

The spatial information in our body model is represented by the image of 11 points as shown in Fig. 2. We then decompose this body model into triplets of points. Since any three non-collinear points in the 3D space define a planar surface, this effectively breaks down the articulated human body into a collection of planar surfaces defined by every non-degenerate triplet in the image plane. Our templates also encode temporal information in the form of pose transition. We thus match a pose pair $\langle I_i, I_j \rangle$ and a template $\langle T_1^k, T_2^k \rangle$ by comparing their pose transitions[1]. Using our point-triplet representation has the following advantages:

- The similarity of pose transitions for articulated bodies can be measured by matching the rigid motions of scene planes defined by all triplets of body points.
- The representation leads to a highly over-determined formulation of the problem, allowing us to achieve robustness to noise and self-occlusions: Given n body point correspondences, we obtain $\binom{n}{3}$ criteria to measure the similarity.
- Anthropometric restrictions can be relaxed, since only the transitions of planes in the 3D space matter, and not the points defining these planes.

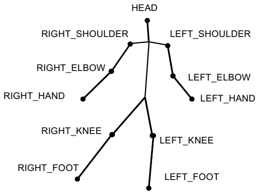


Figure 2. Human body model used in this paper with 11 body points: head, shoulders, elbows, hands, knees and feet.

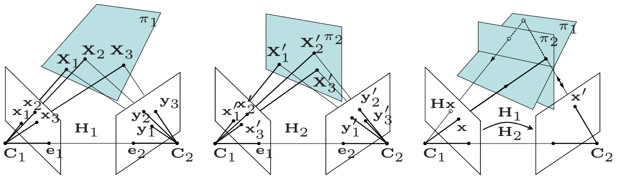

π_1
$\mathbf{X}_1$ $\mathbf{X}_2$ $\mathbf{X}_3$
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$
$\mathbf{y}_1$ $\mathbf{y}_2$ $\mathbf{y}_3$
$\mathbf{H}_1$
$\mathbf{C}_1$ $\mathbf{e}_1$ $\mathbf{e}_2$ $\mathbf{C}_2$
π_2
$\mathbf{X}'_1$ $\mathbf{X}'_2$ $\mathbf{X}'_3$
$\mathbf{x}'_1$ $\mathbf{x}'_2$ $\mathbf{x}'_3$
$\mathbf{y}'_1$ $\mathbf{y}'_2$ $\mathbf{y}'_3$
$\mathbf{H}_2$
$\mathbf{Hx}$
$\mathbf{x}$
$\mathbf{x}'$
(a) (b) (c)

Figure 3. Homographies induced by a moving triplet of points. The motion of a triplet of 3D points $\{\mathbf{X}_i\} \rightarrow \{\mathbf{X}'_i\}$ is observed by two stationary cameras, $\mathbf{C}_1$ and $\mathbf{C}_2$, as $\{\mathbf{x}_i\} \rightarrow \{\mathbf{x}'_i\}$ and $\{\mathbf{y}_i\} \rightarrow \{\mathbf{y}'_i\}$. Together with the epipoles $\mathbf{e}_1 \leftrightarrow \mathbf{e}_2$, the point correspondences induce two homographies $\mathbf{H}_1$ and $\mathbf{H}_2$. For similar motions of triplets, the homography $\mathbf{H} = \mathbf{H}_2^{-1}\mathbf{H}_1$ reduces to a homology with two identical eigenvalues, providing thus a constraint for identifying similar pose transitions.

Homographies Induced by a Triplet of Body Points

Consider the case that $I_{i,j}$ corresponds to $T_{1,2}^k$, and the transformation from I_i to I_j corresponds to that from T_1^k to T_2^k. $I_{i,j}$ and $T_{1,2}^k$ can then be regarded as the images of same moving subject viewed by two different cameras $\mathbf{P}_1$ and $\mathbf{P}_2$, respectively. $\mathbf{P}_1$ and $\mathbf{P}_2$ may have different internal and external parameters. These point correspondences induce an epipolar geometry via the fundamental matrix $\mathbf{F}$. The projection of the camera center of $\mathbf{P}_2$ on $I_{i,j}$ is given by the epipole $\mathbf{e}_1$, which is the right null vector of $\mathbf{F}$. Similarly the image of the camera center of $\mathbf{P}_1$ on $T_{1,2}^k$ is the epipole $\mathbf{e}_2$ given by the right null vector of $\mathbf{F}^T$. Note that this fundamental matrix does not correlate the entire scene, but only the body points.

Let us now consider an arbitrary triplet of 3D body points, $\mathbf{\Delta} = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$, which corresponds to $\Delta_i = \langle \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \rangle$ in I_i and $\Delta_1 = \langle \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3 \rangle$ in T_1^k. After the pose transformation, $\mathbf{\Delta}$ transforms to $\mathbf{\Delta}' = \{\mathbf{X}'_1, \mathbf{X}'_2, \mathbf{X}'_3\}$, which corresponds to $\Delta_j = \langle \mathbf{x}'_1, \mathbf{x}'_2, \mathbf{x}'_3 \rangle$ in I_j and $\Delta_2 = \langle \mathbf{y}'_1, \mathbf{y}'_2, \mathbf{y}'_3 \rangle$ in T_2^k, as illustrated in Fig. 3. $\mathbf{\Delta}$ and $\mathbf{\Delta}'$ determine two scene planes π_1 and π_2 in the 3D space, which induce two homographies $\mathbf{H}_1$ and $\mathbf{H}_2$ between $\mathbf{P}_1$ and $\mathbf{P}_2$. These plane-induced homographies can be computed given four point correspondences, i.e. the image point correspondences $\mathbf{x}_i \leftrightarrow \mathbf{y}_i$ and the epipoles $\mathbf{e}_1 \leftrightarrow \mathbf{e}_2$

A degenerate case occurs when three of the four points are collinear. In general, we can simply discard these degenerate triplets, because the number of non-degenerate triplets exceeds by far the degenerate triplets. A special case is when the epipole is at or close to infinity, all triplets then may be regarded as degenerate. We solve this problem by transforming the image points in projective space in a man-

[1] For brevity of notation, we will hereafter denote a pair $\langle I_i, I_j \rangle$ as $I_{i,j}$.

ner similar to the method described in [19]. The idea is to find a pair of projective transformations corresponding to each image, such that after transformation the epipoles and transformed image points are not at infinity.

Constraints on Homographies due to Moving Triplets

As described above, during a pose transition, the motion of a triplet $\mathbf{\Delta} \rightarrow \mathbf{\Delta}'$ specifies a rigid motion of a scene plane $\pi_1 \rightarrow \pi_2$, which induces two homographies $\mathbf{H}_1$ and $\mathbf{H}_2$ (see Fig. 3). The interesting key observation that leads to our solution is that these homographies define a mapping from I_i (or I_j) to itself given by

$$\mathbf{H} = \mathbf{H}_2^{-1}\mathbf{H}_1.$$

As shown in Fig. 3, $\mathbf{H}$ first maps a point $\mathbf{x}$ on I_i (or I_j) to $\mathbf{y}$ on T_1^k (or T_2^k) through π_1, and then transforms it back to I_i (or I_j) as $\mathbf{Hx}$ through π_2. It can be readily verified either algebraically or from Fig. 3 that points on the intersection of π_1 and π_2 are fixed during the mapping. Another fixed point under this mapping is the epipole $\mathbf{e}_1$. Thus the homography $\mathbf{H}$ has a line of fixed points (the intersection line of π_1 and π_2) and a fixed point not on the line (the epipole $\mathbf{e}_1$), which means that:

Proposition 1 *If a triplet of 3D points observed by two cameras undergo the same motion, then the homography* $\mathbf{H}$ *reduces to a planar homology, and hence two of its eigenvalues must be equal.*

The equality of the two eigenvalues of $\mathbf{H}$ defines a consistency constraint on $\mathbf{H}_1$ and $\mathbf{H}_2$, imposing the assumption that the two cameras are observing the same scene plane motions. In the special case when the triplet is stationary, this equality constraint is still satisfied since $\mathbf{H}$ reduces to an identity matrix, with all its eigenvalues equal to 1. In practice, due to noise and subject-dependent differences, this constraint of equality of two eigenvalues for the same pose transition can be expressed by defining the following error function on $\mathbf{H}$:

$$E(\mathbf{H}) = |a - b|/|a + b|, \qquad (1)$$

where a, and b are the two closest eigenvalues of $\mathbf{H}$. $E(\mathbf{H})$ can be used to measure the similarity of motion of a triplet between two sequences, and the combination of $E(\mathbf{H})$ for all non-degenerate triplets provides a measure of similarity between *pose transitions* $I_i \rightarrow I_j$ and $T_1^k \rightarrow T_2^k$:

$$\mathcal{E}(I_i \rightarrow I_j, T_1^k \rightarrow T_2^k) = \underset{\text{all } \mathbf{\Delta} \rightarrow \mathbf{\Delta}'}{\text{Median}} \left(E(\mathbf{H})\right). \qquad (2)$$

$\mathcal{E}(I_i \rightarrow I_j, T_1^k \rightarrow T_2^k)$ is minimal for similar pose transitions, and is invariant to camera calibration matrix and viewpoint variations.

We now summarize our algorithm to compute the matching score for $I_{i,j}$ and $T_{1,2}^k$ based on pose transition:

1. Suppose that n body point correspondences are given by $\{\mathbf{m}_p^i\}, \{\mathbf{m}_p^j\}, \{\mathbf{m}_p^1\}, \{\mathbf{m}_p^2\}$, for I_i, I_j, T_1^k and T_2^k, respectively, where $p = 1, \ldots, n$. Compute the fundamental matrix $\mathbf{F}$ and epipoles from points correspondences $\mathbf{m}_p^i \leftrightarrow \mathbf{m}_p^1$ and $\mathbf{m}_p^j \leftrightarrow \mathbf{m}_p^2$.
2. For each non-degenerate triplet $\mathbf{\Delta} \rightarrow \mathbf{\Delta}'$, compute the homographies $\mathbf{H}_1$ and $\mathbf{H}_2$ as described above. Then compute the planar homology $\mathbf{H} = \mathbf{H}_2^{-1}\mathbf{H}_1$ and eigenvalues equality error $E(\mathbf{H})$. Combine all non-degenerate triplets to compute $\mathcal{E}(I_i \rightarrow I_j, T_1^k \rightarrow T_2^k)$.
3. The matching score, which may be interpreted as a likelihood function is then given by:

$$\mathcal{S}_{eig}(I_{i,j}, T_{1,2}^k) = 1 - \mathcal{E}(I_i \rightarrow I_j, T_1^k \rightarrow T_2^k) \qquad (3)$$

2.2. Pose recognition from a video sequence

Here we describe our solution for recognizing human poses in a video sequence taken by an uncalibrated camera from arbitrary viewpoints. Suppose we are given a video sequence $\{I_{1\ldots n}\}$, where I_i is the body point representation of the human pose in i^{th} frame, and n is the total number of frames. We are also given the temporal template $T_{1,2}^k$ of some known pose $\mathcal{P}^k$. The procedure for recognizing $\mathcal{P}^k$, if it exists in $\{I_{1\ldots n}\}$, is described as follows:

1. For each frame I_i, $i = 1 \ldots n$, find its d succeeding frames in the video $\{I_{i+1}, I_{i+2}, \ldots, I_{i+d}\}$.
2. The body pose in I_i is recognized as $\mathcal{P}^k$ if

$$\max_j \left\{\mathcal{S}_{eig}(I_{i,j}, T_{1,2}^k) | j = i+1, \ldots, i+d\right\} > \tau,$$

where τ is a threshold.

Checking a segment of d frames allows our method to accommodate for different frame rates of videos and varying execution rates of the motions.

3. Experimental Results

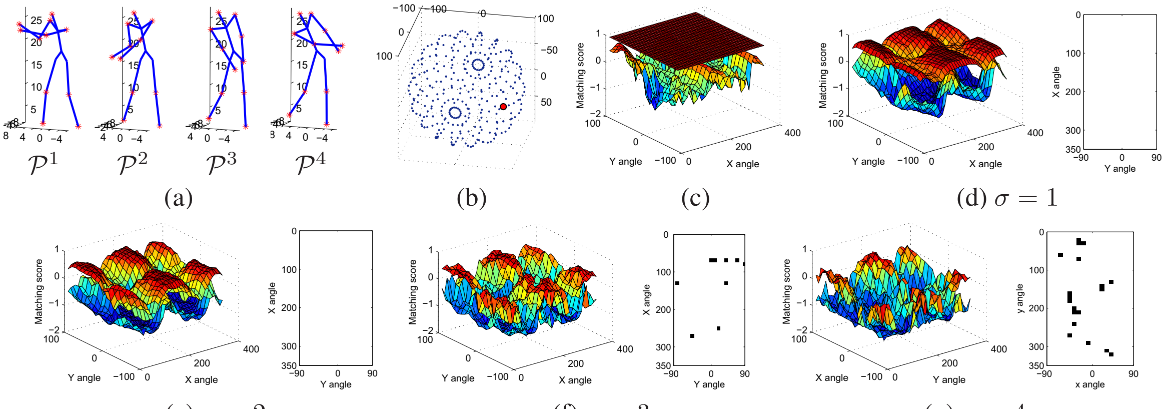
We first present our results on semi-synthetic data, generated from real motion-capture data using synthetic cameras of different internal parameters and viewing directions, with varying noise levels. We then show our test results on real data.

3.1. Evaluation Based on Motion-Capture Data

We generated our semi-synthetic data using the CMU Motion Capture (MoCap) database (http://mocap.cs.cmu.edu/), which consists of recorded 3D points from various real life human actions. The MoCap data were imaged through synthesized cameras with different intrinsic and extrinsic parameters, and Gaussian noise was added. Experiments were then carried out on these semi-synthetic data to evaluate the performance of our method in terms of recognizing poses, in the presence of noise, varying viewpoints, different camera parameters, and subject-dependent variations.

3.1.1 Testing View Invariance

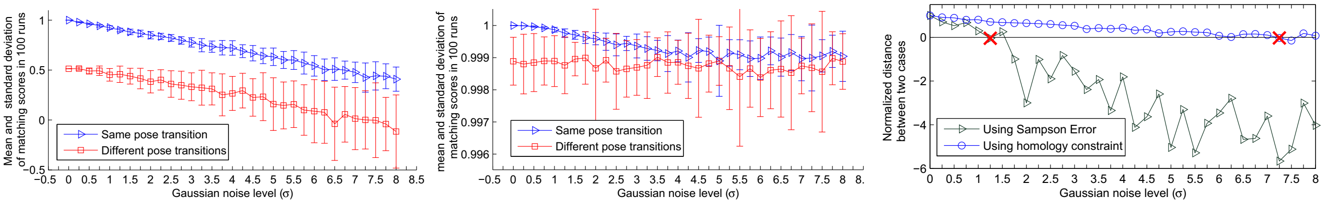
We selected four different poses $\mathcal{P}^1$, $\mathcal{P}^2$, $\mathcal{P}^3$, and $\mathcal{P}^4$ from a golf-swing sequence (see Fig. 4 (a)). These 3D poses are



(a) (b) (c) (d) $\sigma = 1$

(e) $\sigma = 2$ (f) $\sigma = 3$ (g) $\sigma = 4$

Figure 4. View invariance and the matching error for pose transition using a large set of camera orientations with camera locations distributed on a sphere (see text). (a) four 3D poses selected from a golf-swing sequence. (b) the distribution of two cameras - camera 1 is fixed at the location marked in red, and camera 2 is distributed on a sphere around the subject. (c) the plot of two error surfaces for same and different pose transitions. The lower surface corresponds to the error of same pose transitions observed by two cameras for all the configurations shown in (b), while the upper surface corresponds to that of different pose transitions. (d) the plots of error surfaces as in (c) under noise level $\sigma = 1$ on the left, and the black regions on the right show the camera configurations when there is confusion between same and difference pose transition. (e), (f) and (g) show the same plots as (d) under noise levels $\sigma = 2, 3, 4$.



(a) (b) (c)

Figure 6. Results of using our likelihood function against one based on classical Sampson error: (a) and (b) show the plots of matching scores of same and different pose transitions with increasing Gaussian noise for our likelihood function and the Sampson's, respectively. (c) shows the confusion margin in (a) and (b) (see text).

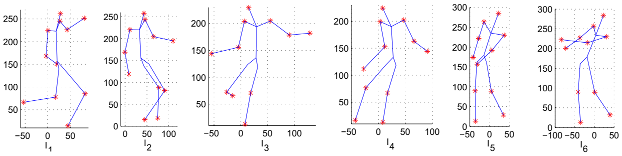


Figure 5. Data used to test robustness to noise. Here, we show the noise-free images observed by two cameras (see text).

observed by two synthesized cameras: camera 1 with focal length of $f_1 = 1000$ looking at the origin of the world coordinate from a fixed location (marked by red color in Fig. 4 (b)), and camera 2 obtained by rotating camera 1 around x and y axes of the world coordinates in increment of 10°, and changing the focal length randomly in the range of 1000 ± 300. Fig. 4 (b) shows all locations of camera 2 as blue points. Let I_1 and I_2 be the images of poses $\mathcal{P}^1$ and $\mathcal{P}^2$ on camera 1, and I_3, I_4, I_5 and I_6 the images of poses $\mathcal{P}^1$, $\mathcal{P}^2$, $\mathcal{P}^3$, and $\mathcal{P}^4$ on camera 2, respectively. Two sets of matching scores were computed for all camera positions: $\mathcal{S}_{eig}(I_{1,2}, I_{3,4})$ and $\mathcal{S}_{eig}(I_{1,2}, I_{5,6})$. The surfaces of matching scores are plotted in Fig.4 (c). The upper flat plane in Fig.4 (c) demonstrates that when two cameras are observing the same pose transition, the error $\mathcal{E}(I_1 \rightarrow I_2, I_3 \rightarrow I_4)$ is always zero in all camera configurations. As shown in this figure, the same pose transition observed by different cameras is readily distinguishable from the case of different transition, regardless of the changes in the viewpoint and camera internal parameters.

3.1.2 Testing Robustness to Noise

We selected two poses $\mathcal{P}^1$ and $\mathcal{P}^2$ from a kick-ball sequence and two poses $\mathcal{P}^3$ and $\mathcal{P}^4$ from a golf-swing sequence. The same experiments of section 3.1.1 were then repeated with camera 1 observing $\mathcal{P}^1$ and $\mathcal{P}^2$ as $I_{1,2}$ and camera 2 observing $\mathcal{P}^1$, $\mathcal{P}^2$, $\mathcal{P}^3$, and $\mathcal{P}^4$ as $I_{3,4}$ and $I_{5,6}$, respectively (see Fig. 5). We then added Gaussian noise to the image points, with σ increasing in the range 0 to 8. As in section 3.1.1, two matching scores $\mathcal{S}_{eig}(I_{1,2}, I_{3,4})$ and $\mathcal{S}_{eig}(I_{1,2}, I_{5,6})$, corresponding to same and different poses, were computed. For each noise level (σ), we repeated the experiment for 100 independent trials and the mean and the standard deviation of both error functions were calculated. Examples of error surfaces and confusion areas (black areas) with $\sigma = 1, 2, 3, 4$ are shown in Fig.4 (d)-(g). Fig. 6

(a) shows the result of configuration in which camera 1 and 2 are related by a 90° rotation around y axis (noise free images are shown in Fig. 5). Using equation (3), the mean of matching scores of two cases (same and different poses) are plotted as curves, and standard deviations as error bars in Fig. 6 (a), which shows that the two cases are unambiguously distinguishable until σ is increased to 7.25. Note that in this experiment the size of the subject is about 250×250 pixels (see 5), which indicates that our method performs extremely well under sever noise. We compared our results to those obtained using a more classical likelihood function based on Sampson's error [7]. The plots are shown in Fig. 6 (b). To compare the results in Fig. 6 (a) and (b), we also computed what we refer to as the *confusion margin* for each likelihood function, which is obtained by computing the distance $d(\sigma)$ between minimum of same pose error bars and maximum of different pose error bars at each noise level σ, and then normalizing it using $\hat{d}(\sigma) = d(\sigma)/d(0)$. If the confusion margin is negative, then the error bars overlap, which indicates confusion in recognizing same and different poses. The curves of both likelihood functions are plotted in Fig. 6 (c), and where they go negative are marked by red crosses. Fig. 6 (c) shows that our likelihood function is more robust than one based on classical Sampson's error.

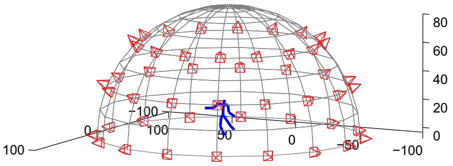


Figure 7. The distribution of cameras used to evaluate view-invariance and camera parameter changes in pose recognition using semi-synthetic data.

3.1.3 Testing pose recognition

We selected 40 poses from the MoCap database: 20 from running motion, 10 from golf swing motion, and 10 from walking motion. Each pose is performed by different actors (6 actors for running, 10 for golf swing, and 6 for walking). For each pose, one actor was selected, and the corresponding pose transition was pictured by a camera with arbitrary focal length and viewpoint, to be used as the template for the pose to be recognized. We thus built a database of templates for the 40 pose transitions, denoted as DB. The rest of the pose instances were used to generate testing data by projecting each instance onto images through 48 cameras distributed on a hemi-sphere, as shown in Fig. 7. We thus built a test dataset of totally $11,520$ 2D pose transitions, with a vast variety of actor behaviors, camera calibrations and viewpoints. Fig.8 shows an example of the same pose observed by some of these cameras. Each 2D pose transition in the test data is matched against each template in DB, and the recognition result is shown in Table 1.

Metric	Value
Tot. Recognition	11,520
True Recognition	10785
True Recog. %	93.62%
Mis-recognition	735
Mis-recogntion %	6.38%

Table 1. Results on testing pose recognition on MoCap data.

3.2. Results on real data

To evaluate our method on real data, we collected a large number of video sequences from Internet, which are taken by cameras with unknown internal parameters and viewpoints. Our data includes video sequences with various actions, such as ballet spin, golf swing, tennis backhand and forehand strokes, tennis serve, etc., and each group includes instances performed by different persons. We selected a number of poses in each action, and built templates for them from one of the instances in that group. When building templates, we set the distance between key pose and successive pose as 5 frames, and during recognition phase we tested using $d = 10$ successive frames for each frame in the video.

Three selected poses and their associated templates are shown in Fig. 9 (a), (c) and (e), where the key poses of the templates are marked in blue and the succeeding poses are marked in red. (a) and (c) are built from a tennis backhand stroke sequence, and their corresponding poses are recognized in two video sequences taken by unknown cameras from different viewpoints. Blue arrows in Fig. 9 (b), (d) and (f) indicate the locations of recognized poses (key poses) and red arrows indicate their succeeding poses. Results on several hundreds of different poses from different action sequences show that by using a single template from an arbitrary viewpoint, our method can recognize poses captured by unknown cameras with different internal parameters and viewpoints.

4. Conclusions

In summary, we propose a novel approach for human body pose recognition by exploiting multi-view geometric constraints in pose transition. Unlike existing methods that study human body pose as a whole, we decompose a body pose into triplets of body points, and compare two pose transitions by matching the motions of their corresponding body point triplets. This reduces the problem from its original non-rigid body pose estimation to that of a better understood and more tractable problem of estimation of the poses of planar surfaces. To this end, we exploit the fact that a moving triplet induces a planar homology between two views, imposing thus an equality constraint on two of its eigenvalues. This constraint is independent of camera calibration matrix and viewpoint, enabling us to achieve view- and camera-invariant recognition of human body poses.

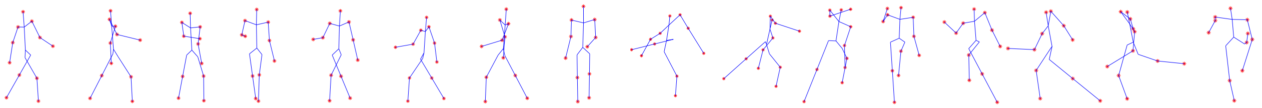
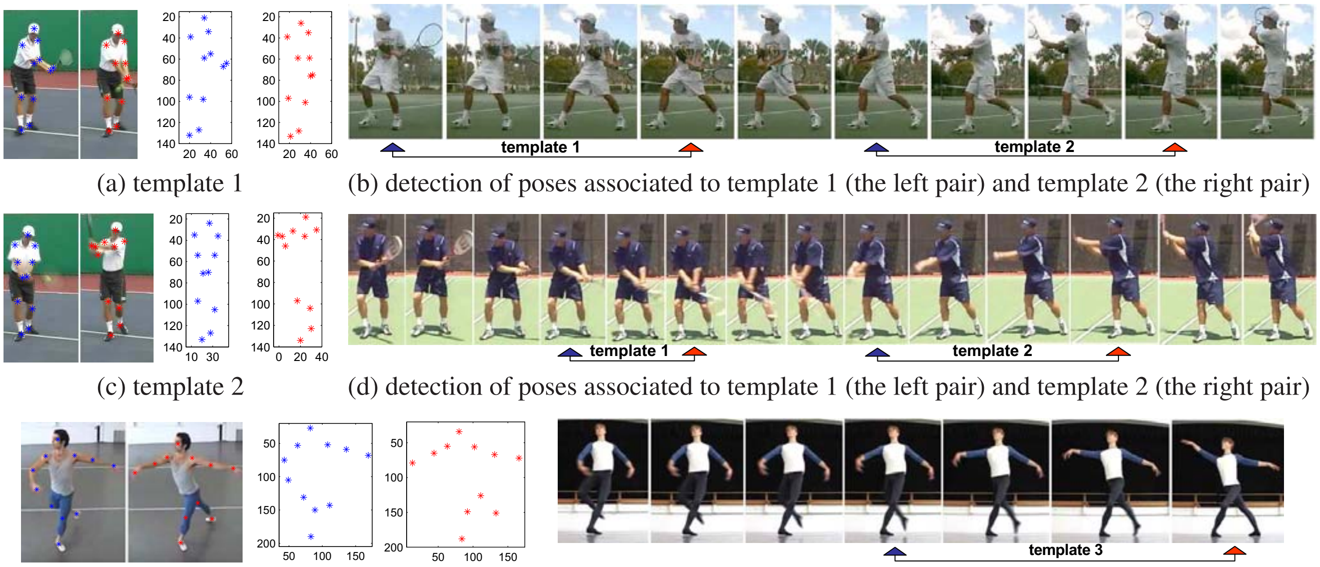


Figure 8. A pose observed from different viewpoints. Note that only 11 body points in red color are used. The stick shapes are shown here for better illustration of pose configuration and extreme variability being handled by our method.



(a) template 1

(b) detection of poses associated to template 1 (the left pair) and template 2 (the right pair)

(c) template 2

(d) detection of poses associated to template 1 (the left pair) and template 2 (the right pair)

(e) template 3

(f) detection of poses associated to template 3

Figure 9. Example results of recognizing human poses in real video sequences collected over Internet.

References

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *CVPR'04*, pages II 882–888, Washington, June 2004. 1

[2] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *TPAMI*, 23(3):257–267, 2001. 1

[3] G. Bradski and J. Davis. Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13(3):174–184, 2002. 1

[4] C. Bregler and J. Malik. Tracking people with twists and exponential maps. *CVPR'98*, pages 8–15, 1998. 1

[5] J. Deutscher and I. Reid. Articulated Body Motion Capture by Stochastic Search. *IJCV*, 61(2):185–205, 2005. 1

[6] M. Dimitrijevic, V. Lepetit, and P. Fua. Human body pose detection using Bayesian spatio-temporal templates. *CVIU*, 104(2-3):127–139, 2006. 1

[7] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 5

[8] N. Howe, M. Leventon, and W. Freeman. Bayesian reconstruction of 3d human motion from single-camera video. *Neural Information Processing Systems*, 1999, 1999. 1

[9] L. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion. *TPAMI*, 22(12):1453–1459, 2000. 1

[10] T. Moeslund and E. Granum. 3D human pose estimation using 2D-data and an alternative phase space representation. *Workshop on Human Modeling, Analysis and Synthesis at CVPR*, 16, 2000. 1

[11] G. Mori and J. Malik. Recovering 3 D Human Body Configurations Using Shape Contexts. *IEEE Trans. PAMI*, 28(7):1052–1062, 2006. 1

[12] E. Ong, A. Micilotta, R. Bowden, and A. Hilton. Viewpoint invariant exemplar-based 3D human tracking. *CVIU*, 104(2-3):178–189, 2006. 1

[13] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. *ECCV*, 2:702–718, 2000. 1

[14] C. Sminchisescu, A. Kanaujia, Z. Li, and D. N. Metaxas. Discriminative density propagation for 3d human motion estimation. In *CVPR*, pages 390–397, 2005. 1

[15] C. Sminchisescu and A. Telea. Human pose estimation from silhouettes. a consistent approach using distance level sets. *WSCG International Conference on Computer Graphics, Visualization and Computer Vision*, 2002. 1

[16] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3D body tracking. *CVPR'01*, 1:447–454, 2001. 1

[17] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3D human tracking. *CVPR'03*, 1, 2003. 1

[18] C. Taylor. Reconstruction of articulated objects from point correspondences ina single uncalibrated image. *CVPR'00*, 1, 2000. 1

[19] Z. Zhang and C. Loop. Estimating the fundamental matrix by transforming image points in projective space. *CVIU*, 82(2):174–180, 2001. 3