# View-invariant Action Recognition from Point Triplets

Yuping Shen, *Student Member, IEEE,* and Hassan Foroosh, *Senior Member, IEEE*

***Abstract*— We propose a new view-invariant measure for action recognition. For this purpose, we introduce the idea that the motion of an articulated body can be decomposed into rigid motions of planes defined by triplets of body points. Using the fact that the homography induced by the motion of a triplet of body points in two identical *pose transitions* reduces to the special case of a homology, we use the equality of two of its eigenvalues as a measure of the similarity of the *pose transitions* between two subjects, observed by different perspective cameras and from different viewpoints. Experimental results show that our method can accurately identify human *pose transitions* and actions even when they include dynamic timeline maps, and are obtained from totally different viewpoints with different unknown camera parameters.**

***Index Terms*— View invariance, homology, pose transition, action recognition, action alignment**

## I. Introduction

Human action recognition has been the subject of extensive studies in the past, highlighted in recent survey papers such as [7], [14], [15], [31]. The main challenges are due to perspective distortions, differences in viewpoints, unknown camera parameters, anthropometric variations, and the large degrees of freedom of articulated bodies [35]. To make the problem more tractable, researchers have made simplifying assumptions on one or more of the following aspects: (1) camera model, such as scaled orthographic [26] or calibrated camera [32]; (2) camera pose, i.e. little or no viewpoint variations; (3) anatomy, such as isometry [16], [17], coplanarity of a subset of body points [16], [17], etc.

There are mainly two lines of research to tackle view-invariance: One is based on the assumption that the actions are viewed by multiple cameras, such as [2], [5], [13], [32], and the second is based on assuming that the actions are captured in monocular sequences by stationary cameras [9], [16], [17], [22], [26], [28], [33], [34]. The obvious limitation of multi-camera approach is that most practical applications are limited to a single camera. In the second category several ideas have been explored, e.g. the invariants associated with a given camera model, such as affine [22], or projective [16], [17], rank constraints on the action space represented by a set of basis functions [26], or the use of epipolar geometry induced by the same pose in two views [9], [28], [33], [34].

The approach proposed in this paper falls in the second category and is based on geometry. We assume a fully projective camera with no restrictions on pose and viewing angles. Moreover, our formulation relaxes restrictive anthropometric assumptions such as isometry. This is due to the fact that unlike existing methods that regard an action as a whole, or as a sequence of individual poses, we represent an action as a set of non-rigid *pose transitions* defined by triplets of points - that is we decompose each pose into a set of point-triplets and find invariants for the rigid motion of the planes defined by triplets across two frames. Therefore, the matching score in our method is based on *pose transition*, instead of being based directly on individual poses or on the entire action. Our approach can also accommodate the possibility of self-occlusion, which may occur under some poses.

This project was in part supported by Electronic Arts - Tiburon.

Y. Shen and H. Foroosh are with Computational Imaging Laboratory (CIL) of the School of EECS at the University of Central Florida (UCF),Orlando, Florida.
Email: {ypshen,foroosh}@cs.ucf.edu

(a) (b)

Fig. 1. Two distinct actions with corresponding poses. (a) The subject holds the same pose in the sequence. (b) same pose as (a) but the subject is also performing a rotation around an axis.

The rest of the paper is organized as follows. In section II, we describe our solution for view-invariant pose and action recognition. Section III presents the experimental results that validate our method on both semi-synthetic and real data. We conclude the paper in section III-C with a discussion of the proposed method and comparison with other methods proposed in the literature.

## II. Action Recognition Based on Homographies Induced by Point Triplets

### A. Representation of Pose

Set of body points is a widely used representation of human pose in action recognition, partly due to the fact that human body can be modeled as an articulate object, and partly because body points can capture sufficient information to achieve the task of action recognition [9], [11], [16], [34]. Using this representation, an action is regarded as a sequence of point sets, or a set of point trajectories in time. Other representations of pose include subject silhouette [3], [4], [28], optical flow [6], [30], [38] and local space time features [12], [24].

In this paper, we use the body point representation. Thus, an action is represented as a sequence of point sets. These points, which are the only inputs to our algorithm, can be obtained by using articulated object tracking techniques such as [18], [20], [21], [23], [27]. Further discussions on articulated object tracking can be found in [1], [7], [14], and is beyond the scope of this paper. We shall, henceforth, assume that tracking has already been performed on the data, and that we are given a set of labeled points for each image.

### B. Pose Transitions

Since action can be regarded as a sequence of poses, a straightforward approach to match two actions is to check the pose-to-pose correspondences. Two same body poses observed by different cameras are related by epipolar geometry via the fundamental matrix, which provides a constraint to match the two poses, regardless of camera calibration matrices or viewpoints. This has motivated the research reported in [9], [26], [28] that are based on the fundamental matrix. Pose-to-pose correspondence, however, is a necessary, but not a sufficient condition for action correspondence. Consider the following case: A subject holds a pose as illustrated in Fig. 1 (a) during the sequence 1, while in sequence 2 (Fig. 1 (b)) it performs a spin, i.e. a rotation around some axis while keeping the same pose as the subject in Fig. 1 (a). These two actions are obviously distinct; however there exist many pose-to-pose correspondences between them since the pose remains unchanged during the two actions. Therefore, additional constraints other than pose correspondence are required to tackle this problem. In addition, most fundamental matrix based methods enforce the constraint that all pose-to-pose correspondences share the same epipolar geometry, i.e., the same fundamental matrix, which is critical to the success of these methods.

Another limitation of fundamental matrix based methods is that they require at least 7 or 8 point correspondences for each pair of poses to measure their similarity. However, in practice, in order to overcome errors, they require far more points, which may not be always possible, especially when self-occlusions exist. For pose-to-pose based methods, this requirement is repeated for every pair of poses (i.e. every image pair), increasing thus their noise sensitivity. We overcome this problem by decomposing body pose into point triplets leading to a largely over-determined problem as described below.

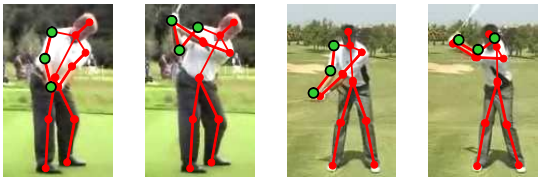
Since actions are spatio-temporal data in 4D, the temporal information is essential to the perception and understanding of actions. However, this is ignored when working directly on individual poses, as in the methods based on fundamental matrix. We alleviate this problem by measuring the similarity between *pose transitions*, rather than poses themselves. Pose transition includes the temporal information of human motion, while keeping the task at the atomic level. Thus, an action can be regarded as a sequence of pose transitions. In the example shown in Fig. 1, although sequences (a) and (b) have the same poses, they are performing different sequences of pose transitions, making it possible to distinguish between the two actions.

*Statement 1:* Two actions are identical if and only if they start at the same pose, and follow the same sequences of pose transitions.

This statement implies that the recognition and matching of two actions can be achieved by measuring the similarity between their sequences of pose transitions. The problem is then reduced to matching pose transitions, which is stated as follows: given two pairs of poses, denoted by $\langle I_1, I_2 \rangle$ and $\langle J_i, J_j \rangle$ (see Fig. 2), we aim to determine whether the transformation from $I_1$ to $I_2$ matches to that from $J_i$ to $J_j$. Note that $I_{1,2}$ and $J_{i,j}$ are sets of 2D labeled points that are observed by cameras with different intrinsic parameters and from different viewpoints.

### C. Matching Pose Transitions

*1) Homographies Induced by a Triplet of Body Points:* Using point representation, a pose is characterized by a set of body points. Each triplet of non-collinear points specifies a scene plane. Therefore, a non-rigid pose transition can be decomposed into the rigid motions of scene planes determined by all non-collinear triplets. This has the following advantages: (1) 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 - rigid motions of planes is a much better understood and more widely studied subject. (2) The representation leads to a highly over-determined formulation of the problem, allowing thus to achieve robustness to noise and self-occlusions: Given $n$ point correspondences, we obtain $\binom{n}{3}$ criteria to measure the similarity. Even if there exist occluded body points, they can be ignored since by far the vast majority of point triplets would be typically available to fulfill the task. (3) Anthropometric restrictions can be relaxed, since only the transitions of planes in the 3D space matter, and not the points defining these planes or the ratios of the distances between these points.



$I_1$ $I_2$ $J_i$ $J_j$

Fig. 2. An example of similar pose transitions. The transition from $I_1$ to $I_2$ is similar to that from $J_i$ to $J_j$. A pose transition can be decomposed into a set of moving triplets of points, e.g., the highlighted triplet of points.

Consider the case that $\langle I_1, I_2 \rangle$ corresponds to $\langle J_i, J_j \rangle$, and the transformation from $I_1$ to $I_2$ corresponds to that from $J_i$ to $J_j$. $I_{1,2}$ and $J_{i,j}$ can then be regarded as the images of same moving subject viewed by two different cameras. Suppose that $I_{1,2}$ are observed by camera $\mathbf{P}_1$ and $J_{i,j}$ by camera $\mathbf{P}_2$. $\mathbf{P}_1$ and $\mathbf{P}_2$ may have different intrinsic and extrinsic parameters. As described earlier, these point correspondences induce an epipolar geometry via the fundamental matrix $\mathbf{F}$. The projection of the camera center of $\mathbf{P}_2$ in $I_1$ and $I_2$ is given by the epipole $\mathbf{e}_1$, which is found as the right null vector of $\mathbf{F}$. Similarly the image of the camera center of $\mathbf{P}_1$ in $J_i$ and $J_j$ is the epipole $\mathbf{e}_2$ given by the right null vector of $\mathbf{F}^T$.

Let us now consider an arbitrary triplet of 3D body points (see highlighted points in Fig. 2 for example), $\mathbf{\Delta} = \{\mathbf{X}_{1,2,3}\}$, which corresponds to $\Delta_1 = \langle \mathbf{x}_{1,2,3} \rangle$ in $I_1$ and $\Delta_i = \langle \mathbf{y}_{1,2,3} \rangle$ in $J_i$. After the pose transformation, $\mathbf{\Delta}$ transforms to $\mathbf{\Delta}' = \{\mathbf{X}'_{1,2,3}\}$, which corresponds to $\Delta_2 = \langle \mathbf{x}'_{1,2,3} \rangle$, in $I_2$ and $\Delta_j = \langle \mathbf{y}'_{1,2,3} \rangle$ in $J_j$, as illustrated in Fig. 3. $\mathbf{\Delta}$ and $\mathbf{\Delta}'$ determine two scene planes $\pi_1$ and $\pi_2$ 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 $\mathbf{e}_1 \leftrightarrow \mathbf{e}_2$:

$$\mathbf{H}_1 \mathbf{x}_i \sim \mathbf{y}_i \quad (1)$$

$$\mathbf{H}_1 \mathbf{e}_1 \sim \mathbf{e}_2 \quad (2)$$

where, as is customary, $\sim$ indicates projective equality up to an unknown scale. A similar set of equations provide $\mathbf{H}_2$:

$$\mathbf{H}_2 \mathbf{x}'_i \sim \mathbf{y}'_i \quad (3)$$

$$\mathbf{H}_2 \mathbf{e}_1 \sim \mathbf{e}_2 \quad (4)$$

Degenerate configurations are discussed later.

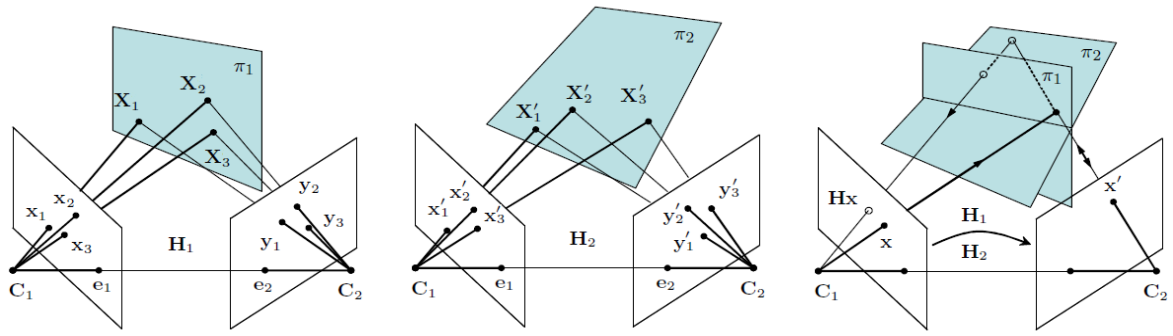



Fig. 3. Homographies induced by a moving triplet of points. Suppose that the motion of a triplet of 3D points $\{\mathbf{X}_i\} \rightarrow \{\mathbf{X}'_i\}$ is observed by two stationary cameras, $\mathbf{C}_1$ and $\mathbf{C}_2$, as $\{\mathbf{x}_i\} \rightarrow \{\mathbf{x}'_i\}$ and $\{\mathbf{y}_i\} \rightarrow \{\mathbf{y}'_i\}$. Together with the epipoles $\mathbf{e}_1 \leftrightarrow \mathbf{e}_2$, the point correspondences $\mathbf{x}_i \leftrightarrow \mathbf{y}_i$ and $\mathbf{x}'_i \leftrightarrow \mathbf{y}'_i$ induce two homographies $\mathbf{H}_1$ and $\mathbf{H}_2$ from the left view to the right view. A homography that maps the left view to itself is then defined as $\mathbf{H} = \mathbf{H}_2^{-1}\mathbf{H}_1$. For similar motions of triplets, this homography is shown to be a homology [8], [19], [36] and hence with two identical eigenvalues, providing thus a constraint for identifying similar pose transitions (see text for more details).

*2) Constraints on Homographies Induced by Moving Triplets:* 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 $\mathbf{H}_2$. These homographies define a mapping from $I_1$ (or $I_2$) to itself given by

$$\mathbf{H} = \mathbf{H}_2^{-1}\mathbf{H}_1. \tag{5}$$

As shown in Fig. 3, $\mathbf{H}$ first maps a point $\mathbf{x}$ on $I_1$ (or $I_2$) to $\mathbf{x}'$ on $J_i$ (or $J_j$) through $\pi_1$, and then transforms it back to $I_1$ (or $I_2$) as $\mathbf{Hx}$ through $\pi_2$. It can be readily verified either algebraically or from Fig. 3 that points on the intersection of $\pi_1$ and $\pi_2$ 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 $\pi_1$ and $\pi_2$) and a fixed point not on the line (the epipole $\mathbf{e}_1$), which means that

*Statement 2:* 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 - we describe this in more detail shortly. In the special case when the triplet is stationary, i.e., $I_1 = I_2$ and $J_i = J_j$, 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 $\mathbf{H}$:

$$E(\mathbf{H}) = \frac{|a-b|}{|a+b|}, \tag{6}$$

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 triplets of non-collinear points provides a measure of similarity between *pose transitions* $I_1 \rightarrow I_2$ and $J_i \rightarrow J_j$:

$$\mathcal{E}(I_1 \rightarrow I_2, J_i \rightarrow J_j) = \underset{\text{all } \Delta_i}{\text{Median}}(E(\mathbf{H})). \tag{7}$$

$\mathcal{E}(I_1 \rightarrow I_2, J_i \rightarrow J_j)$ is minimal for similar pose transitions, and is invariant to camera calibration matrix and viewpoint variations. Here we use median since it is a robust estimator of the mean of a random variable in the presence of possible outliers.

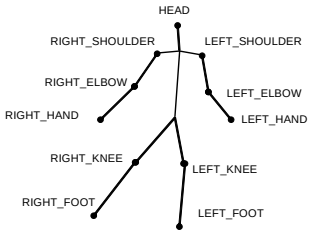



Fig. 4. Human body model used in this paper with 11 body points: head, shoulders, elbows, hands, knees and feet.

### D. Action Alignment and Recognition

The goal of action alignment is to determine the correspondences between two video sequences $A = \{I_{1...n}\}$ and $B = \{J_{1...m}\}$ with matching actions, in our case based on the eigenvalue constraint described above. We align $A$ and $B$ by seeking the optimal mapping $\psi : A \rightarrow B$ such that the cumulative similarity score $\sum_{i=1}^{n} S(i, \psi(i))$ is maximized, where $S(.)$ is the similarity of two poses.

We define $S(.)$ based on matching pose transitions:

$$S(i,j) = \tau - \mathcal{E}(I_{i \rightarrow r_1}, J_{j \rightarrow r_2}), \tag{8}$$

where $\tau$ is a constant threshold, and $r_1, s_1 \in [1, n]$ and $r_2, s_2 \in [1, m]$ are computed as

$$\langle r_1, r_2 \rangle = \underset{r_1, r_2}{\text{argmin}} \{ \min_{s_1, s_2} \mathcal{E}(I_{r_1 \rightarrow s_1}, J_{r_2 \rightarrow s_2}) \}. \tag{9}$$

The matching score of $A$ and $B$ is then defined by

$$\mathscr{S}(A, B) = \max_{\psi} \sum_{i=1}^{n} S(i, \psi(i)). \tag{10}$$

In other words, a pair of reference poses $\langle r_1, r_2 \rangle$ are found first by minimizing (9), and then two pose-transition sequences $A(r_1)$ and $B(r_2)$ are aligned by maximizing (10) using Dynamic Programming (DP). DP has been proven successful in sequence alignment, and has been applied in many areas, such as text processing and bioinformatics. Its application in action recognition can also be found in [16], [22]. The initialization $\langle r_1, r_2 \rangle$ can be further simplified by fixing $r_1$ and $s_1$, e.g., $r_1 = \lfloor \frac{1}{4} n \rfloor$ and $s_1 = \lfloor \frac{3}{4} n \rfloor$.

The traced-back path of DP provides an alignment between two video sequence. Note that this may not be a one-to-one

mapping, since there exists horizontal or vertical lines in the path (see Fig. 10 (c) for example). In addition, due to noise and computational error, different initializations may lead to slightly different valid alignment results. Here the action matching score rather than the alignment is what we are concerned with in action recognition.

**Action recognition :** Consider that we are given a target sequence $\{I_i\}$, and a database of reference sequences corresponding to known different actions, $\{J_i^1\}, \{J_i^2\}, \ldots, \{J_i^K\}$. To recognize the performed action by a subject in $\{I_i\}$, we use the technique in section II-D to align $\{I_i\}$ against all sequences in the database, and recognize it as the action with the highest matching score.

### *E. Degenerate triplets*

A degenerate case occurs when three of the four points in a triplet are collinear. In general, we can simply discard these degenerate triplets, because in practice the number of non-degenerate triplets exceeds by far the degenerate triplets (Note that the total number of available triplets is $\binom{n}{3}$ for $n$ body points).

A special degenerate case is when the epipole is at or close to infinity, in which case all triplets are close to degenerate since the distance between three image points is negligible compared with their distances to the epipole. We solve this problem by transforming the image points in projective space, in a manner similar to Zhang et al. [37]. The idea is to find the projective transformations $\mathbf{P}$ and $\mathbf{P}'$ for each image, such that after transformation the epipoles and transformed image points are not at infinity. Given corresponding image points $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$, we first normalize and transform $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ by $\mathbf{T}_1$ and $\mathbf{T}_2$ respectively such that the RMS distance of the points from the origin is equal to $\sqrt{2}$ and the $x$ and $y$ coordinates of transformed points are $\geq 1$. Then the resulted image points are transformed in projective space by applying the algorithm described in [37].

### *F. Why does this work?*

Any two homographies $\mathbf{H}_1$ and $\mathbf{H}_2$ induced by a pair of scene planes $\pi_1$ and $\pi_2$ can be combined as $\mathbf{H} \sim \mathbf{H}_2^{-1}\mathbf{H}_1$, where $\mathbf{H}$ would always be a homology. An intriguing question that may arise is then the following: If this is true for any two scene planes, then why does the similarity measure based on the eigenvalue constraint proposed above work? and when would this constraint degenerate, i.e. fail to determine that the scene triplets undergo the same motion?

To answer these questions, let us re-examine what we do. A homography induced by a scene plane between two views requires a minimum of four points in order to be specified. We only have three points (i.e. the points of a triplet). However, in our case, the fundamental matrix $\mathbf{F}$ is known - we compute it using all the 11 body points across multiple frames. The key idea that makes it possible to compute the homographies is the fact that all the points on the baseline of the two cameras can be transferred via any scene plane. This is because all the points on the baseline are imaged at the two epipoles, and the epipoles can also be considered as the images of the intersection of the scene plane with the baseline. Therefore, when the fundamental matrix is known, one can use the epipoles as the fourth point for the homography induced by any scene plane. Next using the notations of Fig. 3, the homology $\mathbf{H}$ maps the points $\mathbf{x}_i$ as follows:

$$\mathbf{H}\mathbf{x}_i \sim \mathbf{H}_2^{-1}\mathbf{H}_1\mathbf{x}_i \tag{11}$$

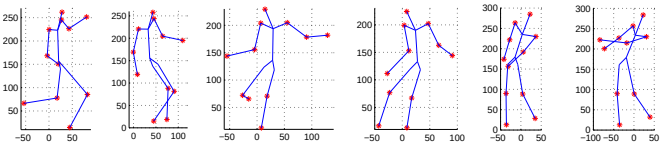
Using equation (3)

$$\mathbf{H}\mathbf{x}_i \sim \mathbf{H}_2^{-1}\mathbf{y}_i \tag{12}$$

By Desargues' theorem [25], lines joining corresponding points must intersect at the vertex of the homology, which in our case is the epipole. This implies that $\mathbf{H}_2^{-1}\mathbf{y}_i$ must lie on the epipolar line of $\mathbf{y}_i$, that is $\mathbf{e}$, $\mathbf{x}_i$, and $\mathbf{H}_2^{-1}\mathbf{y}_i$ must be collinear. This can be expressed as

$$\mathbf{x}_i^T[\mathbf{e}]_\times\mathbf{H}_2^{-1}\mathbf{y}_i = 0 \tag{13}$$

A similar result can be established for $\mathbf{H}_1$. This reveals an interesting result: the homographies induced by a moving triplet must be consistent with the fundamental matrix. This constraint implies that the matrix $\mathbf{H}_i^T\mathbf{F}$ is skew-symmetric [29]. Since we use the epipoles in our computation of the homographies, the consistency condition is satisfied when the points of the triplets viewed in both cameras start and and end in the same positions up to a similarity. However, this is not the only case where the consistency is preserved, since if the points move in such a way that the vertices of the triplet remain along the lines joining the second camera center and the vertices of the triplet, the consistency with the fundamental matrix is still preserved. Fortunately, however, for two different actions it is highly unlikely that this can occur to a body triplet, and even less likely to happen for all possible triplets (e.g. to all the 165 triplets specified by the 11 body points).



$I_1$ $I_2$ $I_3$ $I_4$ $I_5$ $I_6$

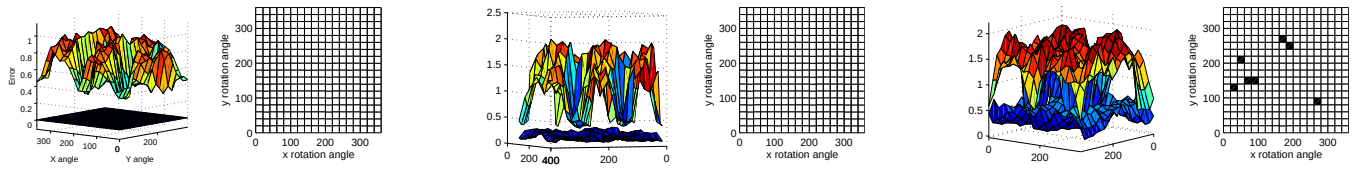
Fig. 5. Data used to test robustness to noise. Here, we show the noise-free images observed by two cameras.

## III. Experimental Results

We validated our approach on both semi-synthetic and real data. We first describe our results on controlled semi-synthetic data, generated from real motion-capture data using synthetic cameras with varying intrinsic parameters, different viewing directions, and varying noise levels. We then show our results on two sets of real data: the IXMAS multiple view data set [32], and our own data set which consists of 56 video sequences of 8 actions (data available at http://cil.cs.ucf.edu/actionrecognition.html).

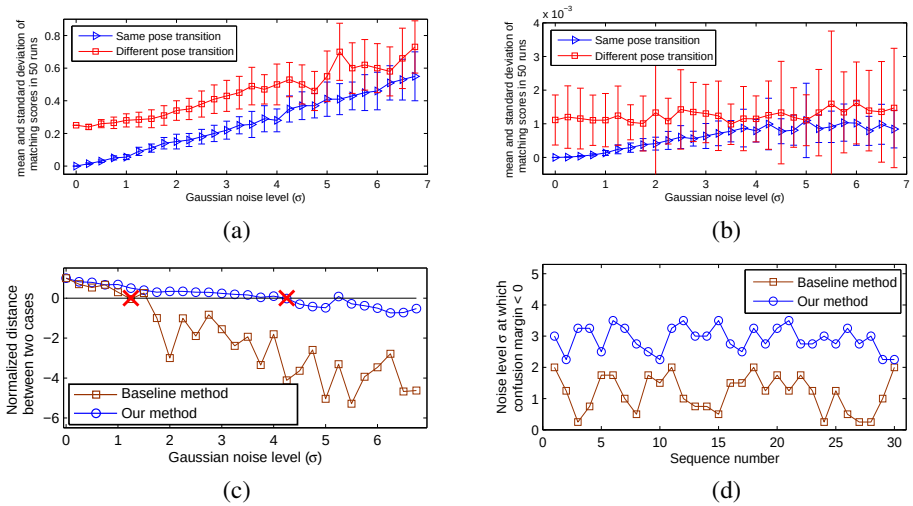
### *A. Analysis Based on Motion-Capture Data*

We generated our semi-synthetic data set using the CMU Motion Capture database (MoCap - http://mocap.cs.cmu.edu/), which consists of sequences of various actions in 3D, captured from real human actions. Here, we do not use the 3D points provided by the data, but merely project the 3D points onto images through synthetic cameras. In other words, we generate the images of 3D body points of a true person, using synthesized cameras and add Gaussian noise. Instead of using all the body points provided in the database, we selected a small subset of body points, which our experiments showed to be sufficient to represent human actions. The body model we employed consists of 11 joints and end points, including head, shoulders, elbows,




(a) $\sigma = 0$ (b) $\sigma = 2$ (c) $\sigma = 4$

Fig. 6. Robustness to noise: for each noise level the plots show the error surfaces for same and different pose transitions. The corresponding grid plots indicate whether there is a confusion: the black blocks in the grid illustrated in plot (c) show the camera orientation angles for which there is confusion between same and different pose transitions. Here camera 2 was obtained by rotating camera 1 around the $x$ and $y$ axes (see text).




(a) (b) (c) (d)

Fig. 7. Performance comparison: (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 error, respectively. Plots correspond to (0, 90) orientation of Fig. 6 for the poses shown in Fig. 5. Plot (c) shows the confusion margin in (a) and (b) (see text); (d) shows the average noise level at which the confusion margin becomes negative for our likelihood function and Sampson error over pose transitions from 30 sequences of different actions: our likelihood function is consistently outperforming Sampson error.

hands, knees and feet (see Fig. 4). Experiments were then carried out on these generated 2D data to evaluate the performance of our method in recognizing pose transitions and actions in the presence of noise, varying viewpoints, different camera parameters, and subject-dependent differences.

*1) Testing View-invariance and Robustness to Noise:* We selected two poses $P_{1,2}$ in a KICK-BALL sequence and two poses $Q_{1,2}$ from the GOLF-SWING sequence (see Fig. 5). These 3D poses are observed by two synthesized cameras: camera 1 with focal length $f_1 = 1000$ looks at the origin of the world coordinate from a fixed location (marked by red color in Fig. 8 (a)), while camera 2 is obtained by rotating camera 1 around $x$ and $y$ axes of the world coordinates in increment of $10^\circ$, and changing the focal length randomly in the range of $1000 \pm 300$. Fig. 8 (a) shows all locations of camera 2 as blue points. Camera 1 observes $P_{1,2}$ as $I_{1,2}$ and camera 2 observes $P_{1,2}$ and $Q_{1,2}$ as $I_{3,4}$ and $I_{5,6}$, respectively (see Fig. 6). We then added Gaussian noise to the image points, with $\sigma$ increasing in steps of 0.25 from 0 to 6.75. Two error functions $\mathcal{E}(I_1 \rightarrow I_2, I_3 \rightarrow I_4)$ and $\mathcal{E}(I_1 \rightarrow I_2, I_5 \rightarrow I_6)$ were computed. For each noise level ($\sigma$), the above procedure was run for 100 independent trials and the mean and the standard deviation of both error functions were calculated. The error surfaces and confusion areas (black areas) with $\sigma = 0, 2, 4$ are shown in Fig.6 (a)-(c). Same and different pose transitions are readily distinguishable up until $\sigma = 4.25$, i.e., up to possibly 12.75 pixel errors. Note that in this experiment the images of the subject have a width of around 150 pixels (see Fig.5), which indicates that our method performs extremely well under sever noise.

We compared our results with those obtained by a baseline method enforcing the equality of fundamental matrices and using the Sampson error [10], [33] (see Appendix A). The plots are shown in Fig. 7. To compare the results in Fig. 7 (a) and (b), we computed what we refer to as the *confusion margin* for each method, 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 $\sigma$, normalized using $\hat{d}(\sigma) = d(\sigma)/d(0)$. If the confusion margin is negative, then the error bars overlap, indicating confusion in recognition. The curves for both methods are plotted in Fig. 7 (c), and where they go negative are marked by red crosses. We repeated the experiments over pose transitions of 30 sequences of different actions. Average noise levels at which confusion occurs for the pose transitions of these 30 sequences are shown in Fig. 7 (d), confirming a superior performance for our method compared with the baseline method.

*2) Testing Action Recognition:* We selected 5 actions from CMU's MoCap data set: walk, jump, golf swing, run, and climb. Each action is performed by 3 actors, and each instance of 3D action is observed by 17 cameras: the first camera was placed on $(x_0, 0, 0)$, looking at the origin of the world coordinate system, while the remaining 16 cameras were generated by rotating around the $y$-axis by $\beta$ and around the $x$-axis by $\alpha$, where $\beta = i\frac{\pi}{4}, i = 0, \ldots, 7$ and $\alpha = j\frac{\pi}{4}, j = 0, 1, 2$ (see Fig. 8 for the location of cameras). The focal lengths were also changed randomly in the range $1000 \pm 300$. Fig. 9 shows an example of a 3D pose observed from 17 viewpoints. We then added Gaussian noise with $\sigma = 3$ to the image points.

Our data set consists of totally 255 video sequences, from

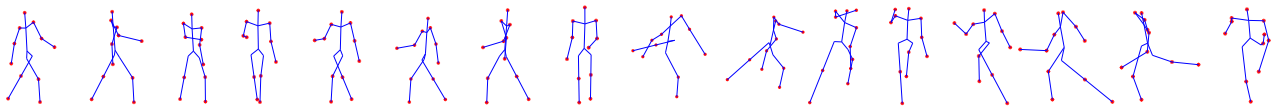


Fig. 9. A pose observed from 17 viewpoints. Note that only 11 body points in red color are used. The stick shapes are shown here for better illustration of pose configurations and for highlighting the extreme variability in appearance that is being handled by our method.

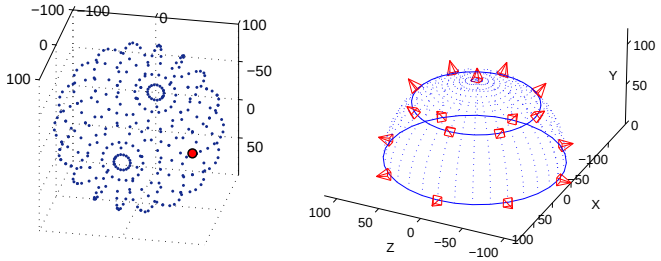



Fig. 8. (a) shows the distribution of two cameras: camera 1 is fixed (red point) while camera 2 is distributed on a sphere around the subject. (b) shows the distribution of cameras used to evaluate view-invariance and camera parameter changes in action recognition using semi-synthetic data.

TABLE I

OUR METHOD: OVERALL ACCURACY ABOUT 92%.

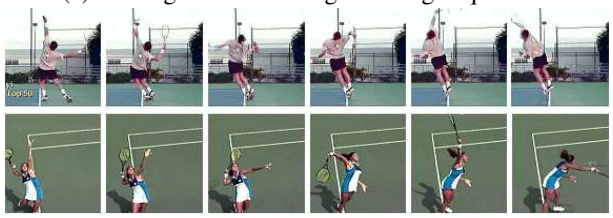
| Ground-truth | Recognized as | | | | |
|---|---|---|---|---|---|
| | Walk | Jump | Golf Swing | Run | Climb |
| Walk | 45 | 1 | | 2 | 2 |
| Jump | 2 | 47 | | 1 | |
| Golf Swing | 1 | | 48 | 1 | |
| Run | | 3 | | 47 | |
| Climb | 6 | 2 | | | 42 |

TABLE II

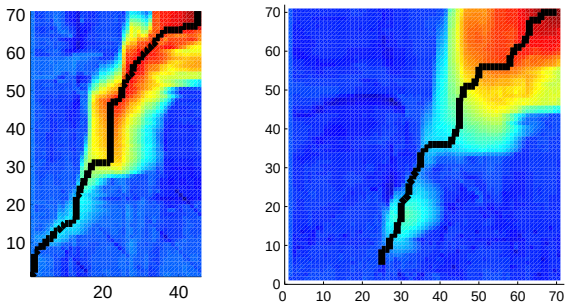
BASELINE METHOD: OVERALL ACCURACY ABOUT 85%.

| Ground-truth | Recognized as | | | | |
|---|---|---|---|---|---|
| | Walk | Jump | Golf Swing | Run | Climb |
| Walk | 41 | 2 | 1 | 4 | 2 |
| Jump | 2 | 45 | | 1 | 2 |
| Golf Swing | 2 | 1 | 45 | 2 | |
| Run | 3 | 2 | 1 | 43 | 1 |
| Climb | 5 | 1 | 1 | 3 | 40 |

(a) An alignment of two golf swing sequences.



(b) An alignment of two tennis-serve sequences.




(c) $\tau = 0.3$ (d) $\tau = 0.4$

Fig. 10. Two examples of action alignment: (a) shows the frame-by-frame mappings between the two golf-swing sequences with different lengths, (b) alignment for a tennis-serve action with different starting and ending frames, (c) and (d) show the optimized traced paths using dynamic time warping.

which we generated a reference Action Database (ADB) of 5 sequences, one sequence for each action. These sequences are all selected from viewpoint 1. The rest of the data set was used as test data, and each sequence was matched against all actions in the ADB and classified as the one with highest score. For each sequence matching, 10 random initialization are tested. The classification results are shown in Table I. For instance, the number in row 1, column 5 means that two of walking sequences are misclassified as climbing. Table II shows the confusion matrix for the same data set using the baseline method. The overall classification accuracy for our method is 92%, compared with 85% for the baseline method.

A further examination of the experiments on viewpoint changes is shown in Table III, from which we find that the accuracy for viewpoint 17 in our method is as low as 46.7%. This is most probably due to the severe distortions caused from viewpoint 17, which is directly above the subject - also reported problematic by [16]. Ignoring this highly unlikely viewpoint, the average accuracy for other viewpoints is about 95%, which is remarkably good, despite the extreme viewpoint changes and variations in camera intrinsic parameters.

### B. Results on Real Data

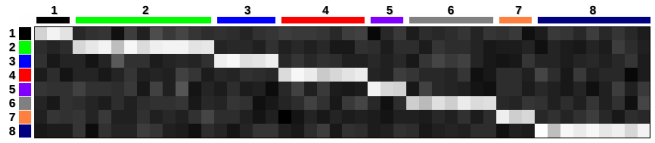
*1) Our own data set:* We evaluated our method on a data set of real video sequences. To best simulate the situations in real life, we collected these videos from the Internet, coming from a variety of sources. The collected data set consists of 56 sequences of 8 actions: 4 of ballet fouettes, 12 of ballet spin, 6 of push-up exercise, 8 for golf swing, 4 of one-handed tennis backhand stroke, 8 of two-handed tennis backhand stroke, 4 of tennis forehand stroke, and 10 of tennis serve. Each action is performed by different subjects, and the videos are taken by different unknown cameras from various viewpoints collected over the Internet. In addition, videos in the same group (action) may have different starting and ending times, thus may be only partially overlapped. Subjects also perform the same action in different ways and at different speeds. Self-occlusion and minor camera motions also exist in many of the sequences, which provide a good test of the robustness of our method.

*1.1) Aligning Two Actions*

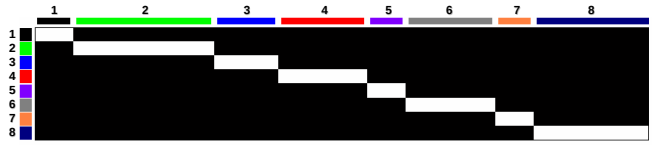
We tested our action alignment approach for numerous sequences in our database, two of which are shown in Fig. 10. These test sequences had different lengths or different starting and

TABLE III
RECOGNITION ACCURACY FOR 17 DIFFERENT VIEWPOINTS

| | Viewpoints | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| # of sequences | 10 | 15 | 15 | 15 | 15 | 15 | 15 | 15 | 15 | 15 | 15 | 15 | 15 | 15 | 15 | 15 | 15 |
| # of errors | 0 | 0 | 1 | 1 | 0 | 1 | 2 | 0 | 2 | 1 | 1 | 0 | 2 | 1 | 1 | 0 | 8 |
| Accuracy | 1.0 | 1.0 | .933 | .933 | 1.0 | .933 | .867 | 1.0 | .867 | .933 | .933 | 1.0 | .933 | .867 | .933 | 1.0 | .467 |



(a) Confusion matrix



(b) Recognition result

Fig. 11. Result of action recognition based on the confusion matrix. The actions are denoted by numbers: 1 - ballet fouette, 2 - ballet spin, 3 - push up, 4 - golf swing, 5 - one-handed tennis backhand stroke, 6 - two-handed tennis backhand stroke, 7 - tennis forehand stroke, 8 - tennis serve. All actions are correctly classified.

ending points of action. Fig. 10 (a) and (b) show the two examples of aligned sequences. In the first example, two sequences of different lengths (the length of the upper sequence is 46 frames, and the lower one is 71 frames) are aligned, in which two players are performing golf swing at different speeds. The alignment result is shown in Fig. 10 (a): the first two rows show some matched poses, and the frame-to-frame mapping of two sequences are displayed in the third row. In the second example, shown in Fig. 10 (b), two sequences of a tenis serve-actions are aligned: the two sequences are roughly of the same length but different start and ending frames in terms of player's pose.

The accumulation matrices and the back-tracked paths in dynamic time warping for these two examples are shown in Fig. 10 (c) and (d), respectively. The thresholds used in these examples were $\tau = 0.3$ and $\tau = 0.4$. The choice of $\tau$ reflects our estimate of the matching noise. The examples with different $\tau$ values shown here are only for demonstration purposes. We found that dynamic time warping in general performs well for a large range of values of $\tau$ and provides good-enough solutions for action recognition using our method. As stated in section II-D, we set $\tau$ to a constant value of 0.3 throughout our experiments.

*1.2) Results on Action Recognition*

We built an action database (ADB) by selecting one sequence for each action. The other sequences were used as test sequences, and were matched against all actions in ADB. The recognition result is based on the highest matching score for each sequence. We show the confusion matrix in Fig. 11 (a), where light colors represent similar actions and dark colors represent dissimilar actions. The recognition result is shown in Fig. 11 (b) (100% match), where the black block in each column indicates the recognized action for each test sequence.

*2) IXMAS Data Set:* We tested our methods on IXMAS data set [32], which is a multiple view data set with 13 daily-live motions, each performed 3 times by 11 actors. We segmented all sequences into different actions based on the provided ground-truth segmentation, and tested on actions {1 2 3 4 5 8 9 10 11 12}, and similar to [32] we applied our method on all actors except for "Pao" and "Srikumar". We took the actor "andreas" out of the data set, and used "andreas 1" under camera view "cam1" as the reference for all actions. The remaining sequences performed by all other actors were used as test data. Note that our reference and test sequences are all single-view, thus multiple view information in the data set is not used in our experiments. The global threshold $\tau = 0.3$ is used in our experiments, and the recognition results are shown in Table IV. The average recognition rate is 90.23%, which is comparable to the result of 93.33% using MHV [32] given that we do not use multiple images and rely only on one view.

TABLE IV
RECOGNITION RATE FOR IXMAS DATA. ACTIONS ARE: 1= CHECK WATCH, 2=CROSS ARMS, 3=SCRATCH HEAD, 4=SIT DOWN, 5= GET UP, 8=WAVE, 9= PUNCH, 10= KICK, 11=POINT, 12=PICK UP

| Action | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Recognition Rate % | 89.6 | 94.8 | 85.2 | 91.9 | 91.1 |
| Action | 8 | 9 | 10 | 11 | 12 |
| Recognition Rate % | 85.2 | 92.6 | 91.9 | 90.4 | 89.6 |

TABLE V
COMPARISON OF DIFFERENT METHODS.

| Methods | # of views | Camera model | Input | Other assumptions |
|---|---|---|---|---|
| Ours | 1 | Persp. projective | Body points | |
| [16] | 1 | Persp. projective | Body points | five preselected coplanar points or limbs trace planar area |
| [32] | 5 | Persp. projective | Visual hulls | |
| [28] | 1 | Persp. projective | Feature points | |
| [2] | $> 1$ | Persp. projective | Optical flow and silhouettes | |
| [22] | 1 | Affine | Body points | |
| [33] | 1 | Affine | Silhouettes | |
| [26] | 1 | Affine | Body points | Same start and end of sequences |
| [9] | 1 | Persp. projective | Body points | |

## C. Discussion and Conclusion

Table V summarizes the existing methods for view-invariant action recognition. In regards to the number of required cameras, the existing methods fall into two groups: multiple view methods ( [16], [2], etc.) and monocular methods ( [9], [16], [22], [26], [28], [33] and ours). Multiple view methods are more expensive, and are restricted in most real life problems when only one camera is available, e.g. monocular surveillance. Some methods are based on the assumption of affine camera model ( [22], [26], [33]), which can be readily violated in many practical

applications. Most pose-to-pose human action recognition techniques assume also some anthropometric restrictions, such as isometry. Some methods also make additional assumptions. For instance, Parameswaran et al. [16], assume that canonical poses are predefined in the input data, or certain limbs trace areas on a plane during actions; Sheikh et al. [26] assume that each action is a span of a number of action bases, which is defined by a number of example sequences in the same action. This implicitly implies that the start and the end of input sequences are restricted to those of example sequences. In addition, a large number of examples may be required, due to the irregularities of human action.

Assuming that a set of 11 points are tracked, our approach makes no special assumptions about the camera internal and external geometry or the human body, and yet as demonstrated on both semi-synthetic and real data, we are capable of tackling variations in camera parameters, extreme changes in the camera viewpoints across different subjects, and high noise level in data points.

### *A. Appendix: Template matching based on fundamental matrix between views*

When a pose transition $I_1 \rightarrow I_2$ matches $J_1 \rightarrow J_2$, one can regard the corresponding pairs (e.g. $I_1$ and $J_1$) as perspective views of the same rigid object. Since two perspective views of a rigid object are related via epipolar geometry and their associated fundamental matrix [10], a straightforward solution to measure the similarity between $I_{1,2}$ and $J_{1,2}$ is to check the equality of the fundamental matrix for all corresponding poses. Suppose we have $2n$ point correspondences $\mathbf{x}_i \leftrightarrow \mathbf{y}_i$, where $I_1 = \{\mathbf{x}_{1...n}\}, I_2 = \{\mathbf{x}_{(n+1),...,2n}\}, J_1 = \{\mathbf{y}_{1...n}\}$, $J_2 = \{\mathbf{y}_{(n+1),...,2n}\}$, and $n$ is the number of body points in each image. A classical error function, based on the fundamental matrix, can be defined as:

$$\mathcal{E}_{fund}(I_{1,2}, J_{1,2}) = \underset{i=1..2n}{\text{Median}} \left\{ \frac{(\mathbf{y_i^T F x}_i)^2}{(\mathbf{Fx}_i)_1^2 + (\mathbf{Fx}_i)_2^2 + (\mathbf{F^T y}_i)_1^2 + (\mathbf{F^T y}_i)_2^2} \right\}, \quad (14)$$

where $(\cdot)_j$ refers to the $j^{th}$ entry of the vector. The second term in (14) is essentially the Sampson cost function [10], which is widely used in estimation methods in multiple view geometry. A scoring function $\mathcal{S}_{fund}(\cdot)$ can then be defined similar to (8), which we used as a baseline method for comparison in the experiments section.

## REFERENCES


[1] J. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

[2] M. Ahmad and S. Lee. HMM-based Human Action Recognition Using Multiview Image Sequences. *ICPR*, pages 263–266, 2006.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proc. ICCV*, volume 2, pages 1395– 1402, 2005.

[4] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[5] F. Cuzzolin. Using Bilinear Models for View-invariant Action and Identity Recognition. *Proc. IEEE CVPR*, 2:1701–1708, 2006.

[6] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *Proc. IEEE International Conference on Computer Vision*, pages 726–733, 2003.

[7] D. Gavrila. Visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[8] L. V. Gool and A. Proesmans. Grouping and invariants using planar homologies.

[9] A. Gritai, Y. Sheikh, and M. Shah. On the use of anthropometry in the invariant analysis of human actions. *Proc. International Conference on Pattern Recognition*, 2, 2004.

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

[11] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.

[12] I. Laptev, S. Belongie, P. Perez, J. Wills, C. universitaire de Beaulieu, and U. San Diego. Periodic Motion Detection and Segmentation via Approximate Sequence Alignment. *Proc. IEEE International Conference on Computer Vision*, pages 816–823, 2005.

[13] F. Lv and R. Nevatia. Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching. *CVPR*, pages 1–8, 2007.

[14] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[15] T. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, 2006.

[16] V. Parameswaran and R. Chellappa. View invariants for human action recognition. *Proc. IEEE CVPR*, 2, 2003.

[17] V. Parameswaran and R. Chellappa. View Invariance for Human Action Recognition. *IJCV*, 66(1):83–101, 2006.

[18] V. Pavlovic, J. Rehg, T. Cham, and K. Murphy. A Dynamic Bayesian Network Approach to Figure Tracking using Learned Dynamic Models. *Proc. ICCV*, 1:94–101, 1999.

[19] P. Pritchett and A. Zisserman. Matching and reconstruction from widely separated views. In *SMILE'98: Proceedings of the European Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, pages 78–92. Springer-Verlag, 1998.

[20] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proc. CVPR, Volume 1*, pages 271–278, 2005.

[21] D. Ramanan, D. Forsyth, and A. Zisserman. Tracking people and recognizing their activities. In *Video Proceeedings of Computer Vision and Pattern Recognition (VPCVPR)*, page 1194, 2005.

[22] C. Rao, A. Yilmaz, and M. Shah. View-Invariant Representation and Recognition of Actions. *International Journal of Computer Vision*, 50(2):203–226, 2002.

[23] J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. *Proc. Int. Conf. Computer Vision*, pages 612–617, 1995.

[24] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. *Proc. International Conference on Pattern Recognition*, 3, 2004.

[25] J. G. Semple and G. Kneebone. *Algebraic Projective Geometry*. Oxford University Press, 1979.

[26] Y. Sheikh and M. Shah. Exploring the Space of a Human Action. *Proc. IEEE International Conference on Computer Vision,*, 1, 2005.

[27] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *Proc. ECCV*, pages 629–644, 2002.

[28] T. Syeda-Mahmood, A. Vasilescu, S. Sethi, I. Center, and C. San Jose. Recognizing action events from multiple viewpoints. *Proc. IEEE Workshop on Detection and Recognition of Events in Video*, pages 64–72, 2001.

[29] T. Viéville and Q.-T. Luong. Motion of points and lines in the uncalibrated case. Technical report 2054, INRIA, 1993.

[30] L. Wang. Abnormal Walking Gait Analysis Using Silhouette-Masked Flow Histograms. *Proc. International Conference on Pattern Recognition*, pages 473–476, 2006.

[31] L. Wang, W. Hu, T. Tan, et al. Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, 2003.

[32] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2-3):249–257, 2006.

[33] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. *Proc. IEEE Computer Vision and Pattern Recognition,*, 1, 2005.

[34] A. Yilmaz and M. Shah. Matching actions in presence of camera motion. *Computer Vision and Image Understanding*, 104(2-3):221–231, 2006.

[35] V. Zatsiorsky. *Kinematics of Human Motion*. Human Kinetics, 2002.

[36] L. Zelnik-Manor and M. Irani. Multiview constraints on homographies. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(2):214–223, 2002.

[37] Z. Zhang and C. Loop. Estimating the fundamental matrix by transforming image points in projective space. *Computer Vision and Image Understanding*, 82(2):174–180, 2001.

[38] G. Zhu, C. Xu, W. Gao, and Q. Huang. Action Recognition in Broadcast Tennis Video Using Optical Flow and Support Vector Machine. *Lecture Notes in Computer Science*, 3979:89, 2006.