View-Invariant Action Recognition Using Fundamental Ratios

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

Abstract

A moving plane observed by a fixed camera induces a fundamental matrix $\mathbf{F}$ *across multiple frames, where the ratios among the elements in the upper left* 2×2 *submatrix are herein referred to as the* Fundamental Ratios. *We show that fundamental ratios are invariant to camera parameters, and hence can be used to identify similar plane motions from varying viewpoints. For action recognition, we decompose a body posture into a set of point triplets (planes). The similarity between two actions is then determined by the motion of point triplets and hence by their associated fundamental ratios, providing thus view-invariant recognition of actions. Results evaluated over 255 semi-synthetic video data with 100 independent trials at a wide range of noise levels, and also on 56 real videos of 8 different classes of actions, confirm that our method can recognize actions under substantial amount of noise, even when they have dynamic timeline maps, and the viewpoints and camera parameters are unknown and totally different.*

1. View-invariant Action Recognition

Human action recognition has been the subject of extensive studies in the past, highlighted in recent survey papers such as [5, 10, 15]. 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 [18]. 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 [13] or calibrated camera [16]; (2) camera pose, i.e. little or no viewpoint variations; (3) anatomy, such as isometry [11], coplanarity of a subset of body points [11], etc.

There are mainly two lines of research to tackle view-invariance: One is based on using multiple cameras, such as [16, 1, 9, 4], and the second is based on multiple frames of a stationary camera. 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, e.g. affine [12], or projective [11], rank constraints on the action space represented by a set of basis functions [13], or the use of epipolar geometry induced by the same pose in different views [14, 17, 6].

1.1. Our Approach: Overview

Our approach falls in the last category. We assume a fully projective unknown camera with no restrictions on pose or viewpoint. Moreover, our formulation relaxes restrictive anthropometric assumptions such as isometry. 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 *pose transitions* defined by all possible triplets of body points, i.e., we break down further each pose into a set of point-triplets and find invariants for the motion of these triplets across frames. Therefore, the matching score in our method is based on *pose transitions* of all possible triplets of body points, instead of being based directly on individual poses or on the entire action.

2. Fundamental Ratios

In this section, we establish specific relations between homographies induced by world planes (determined by any triplet of non-collinear 3D points) and the fundamental matrix associated with two views. More specifically, we derive a set of feature ratios that are invariant to camera intrinsic parameters for a natural perspective camera model of zero skew and unit aspect ratio. We then show that these feature ratios are projectively invariant to similarity transformations of the triplet of points in the 3D space, or equivalently invariant to rigid transformations of camera.

Proposition 1 *Given two cameras* $\mathbf{P}_i \sim \mathbf{K}_i[\mathbf{R}_i|\mathbf{t}_i]$, $\mathbf{P}_j \sim \mathbf{K}_j[\mathbf{R}_j|\mathbf{t}_j]$ *with zero skew and unit aspect ratio, denote the relative translation and rotation from* $\mathbf{P}_i$ *to* $\mathbf{P}_j$ *as* $\mathbf{t}$ *and* $\mathbf{R}$ *respectively, then the upper left* 2×2 *submatrix of the fundamental matrix between two views is of the form*

$$\mathbf{F}^{2\times 2} \sim \begin{bmatrix} \epsilon_{1st}\mathbf{t}^s\mathbf{r}_1^t & \epsilon_{1st}\mathbf{t}^s\mathbf{r}_2^t \\ \epsilon_{2st}\mathbf{t}^s\mathbf{r}_1^t & \epsilon_{2st}\mathbf{t}^s\mathbf{r}_2^t \end{bmatrix}, \quad (1)$$

where $\mathbf{r}_k$ *is the* k*-th column of* $\mathbf{R}$*, the superscript,* e.g. i*, refers to* i^{th} *element of a vector, and* ϵ_{rst} *for* $r,s,t = 1,\ldots,3$ *is a permutation tensor*[1].

Remark 1 *The ratios among elements of* $\mathbf{F}^{2\times 2}$ *are invariant to camera calibration matrices* $\mathbf{K}_i$ *and* $\mathbf{K}_j$.

The upper left 2×2 sub-matrices $\mathbf{F}^{2\times 2}$ for two moving cameras could be used to measure the similarity of camera motions [3]. That is, if two cameras perform the same motion (same relative translation and rotation during the motion), and $\mathbf{F}_1$ and $\mathbf{F}_2$ are the fundamental matrices between

[1] The use of tensor notation is explained in details in [8], p563.

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

any pair of corresponding frames, then $\mathbf{F}_1^{2\times2} \sim \mathbf{F}_2^{2\times2}$. This also holds for the dual problem when the two cameras are fixed, but the scene objects in both cameras perform the same motion. A special case of this problem is when the scene objects are planar surfaces, which is discussed below.

Proposition 2 *Suppose two fixed cameras are looking at two moving planar surfaces, respectively. Let* $\mathbf{F}_1$ *and* $\mathbf{F}_2$ *be the two fundamental matrices induced by the two moving planar surfaces (e.g. by the two triplets of points). If the motion of the two planar surfaces is similar (differ at most by a similarity transformation), then*

$$\mathbf{F}_1^{2\times2} \sim \mathbf{F}_2^{2\times2} \tag{2}$$

where the projective equality, denoted by $\sim$*, is invariant to camera orientation.*

Here similar motion implies that plane normals undergo same motion up to a similarity transformation. The projective nature of the view-invariant equation in (2) implies that the elements in the sub-matrices on the both sides of (2) are equal up to an arbitrary non-zero scale factor, and hence only the ratios among them matter. We call these ratios the *fundamental ratios*, and as propositions 1 and 2 imply, these fundamental ratios are invariant to camera intrinsic parameters and viewpoint. To eliminate the scale factor, we can normalize both sides using $\hat{\mathbf{F}}_i = |\mathbf{F}_i^{2\times2}|/\|\mathbf{F}_i^{2\times2}\|_F, i = 1, 2$, where $|\cdot|$ refers to absolute value operator and $\|\cdot\|_F$ stands for the Frobenius norm. We then have

$$\hat{\mathbf{F}}_1 = \hat{\mathbf{F}}_2 \tag{3}$$

In practice, $\hat{\mathbf{F}}_1$ and $\hat{\mathbf{F}}_2$ may not be exactly equal due to noise, computational errors or subjects' different ways of performing same actions. We, therefore, define the following function to measure the residual error:

$$\mathcal{E}(\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2) = \|\hat{\mathbf{F}}_1 - \hat{\mathbf{F}}_2\|_F \tag{4}$$

3. Action Recognition Using Fundamental Ratios

We are given a video sequence $\{I_t\}$ and a database of reference sequences corresponding to K different known actions, $DB = \{J_t^1\}, \{J_t^2\}, \ldots, \{J_t^K\}$, where I_t and J_t^k are labeled body points in frame t. Our goal is to identify the sequence $\{J_t^k\}$ from DB such that the subject in $\{I_t\}$ performs the closest action to that observed in $\{J_t^k\}$.

Existing methods for action recognition such as [2, 17] consider an action as a whole, which usually requires known start and end frames and is limited when action execution rate varies. Some other approaches such as [6] regard action as a sequence of individual poses, and rely on pose-to-pose similarity measures. Since an action consists of spatio-temporal data, the temporal information plays a crucial role in recognizing action, which is ignored in a pose-to-pose approach. We thus propose using *pose transition*. One can thus compare actions by comparing their pose transitions.

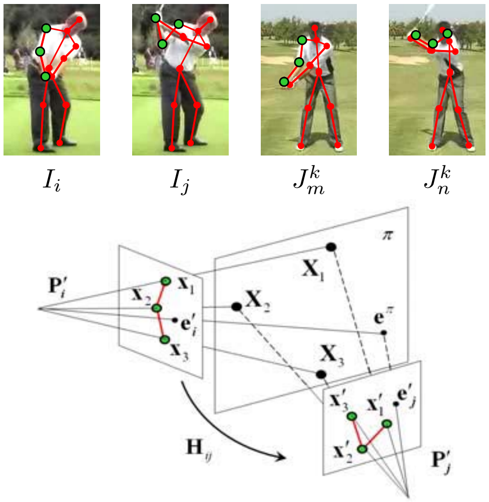


Figure 1. Fundamental matrix induced by a moving plane is dual to a stationary plane with moving camera.

3.1. Matching Pose Transition

The 3D body structure of a human can be divided into triplets of body points (see Fig. 1), each of which determines a plane in the 3D space when the points are not collinear. The problem of comparing articulated motions of human body thus transforms to comparing rigid motions of body planes (triplets). According to proposition 2, the motion of a plane induces a fundamental matrix, which can be identified by its associated fundamental ratios. If two pose transitions are identical, their corresponding body point triplets have the same fundamental ratios, which provide a measure for matching two pose transitions.

3.1.1 Fundamental matrix induced by a moving triplet

We are given an observed pose transition $I_i \rightarrow I_j$ from sequence $\{I_t\}$, and a second one $J_m^k \rightarrow J_n^k$ from sequence $\{J_t^k\}$. When $I_i \rightarrow I_j$ corresponds to $J_m^k \rightarrow J_n^k$, one can regard them as observations of the same 3D pose transition from two different cameras $\mathbf{P}_1$ and $\mathbf{P}_2$, respectively. There are two instances of epipolar geometry associated with this scenario:

1- The mapping between the image pair $\langle I_i, I_j\rangle$ and the image pair $\langle J_m^k, J_n^k\rangle$ is determined by the fundamental matrix $\mathbf{F}$ [8] related to $\mathbf{P}_1$ and $\mathbf{P}_2$. The projection of the camera center of $\mathbf{P}_2$ in I_i or I_j 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_m^k or J_n^k is the epipole $\mathbf{e}_2$ given by the right null vector of $\mathbf{F}^T$.

2- The other instance of epipolar geometry is between transitioned poses of a triplet of body points in two frames of the same camera, i.e. the fundamental matrix induced by a moving body point-triplet, which we denote as $\mathcal{F}$. We call this fundamental matrix the *inter-pose fundamental matrix*, as it is induced by the transition of body point poses viewed by a stationary camera.

Let Δ be a triplet of non-collinear 3D points, whose motion lead to different projections on I_i, I_j, J_m^k and J_n^k as $\Delta_i, \Delta_j, \Delta_m^k$ and Δ_n^k, respectively:

$$\Delta_i = \langle \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \rangle, \Delta_j = \langle \mathbf{x}'_1, \mathbf{x}'_2, \mathbf{x}'_3 \rangle,$$
$$\Delta_m^k = \langle \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3 \rangle, \Delta_n^k = \langle \mathbf{y}'_1, \mathbf{y}'_2, \mathbf{y}'_3 \rangle.$$

Δ_i and Δ_j can be regarded as projections of a stationary 3D point triplet $\langle \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3 \rangle$ on two virtual cameras $\mathbf{P}'_i$ and $\mathbf{P}'_j$, as shown in Fig. 1. $\langle \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3 \rangle$ defines a world plane π, which induces a homography $\mathbf{H}_{ij}$ between $\mathbf{P}'_i$ and $\mathbf{P}'_j$. It is known that a homography may be computed from four corresponding image points. In this case, the four points can be the image points $\mathbf{x}_1, ..., \mathbf{x}_3$ and $\mathbf{x}'_1, ..., \mathbf{x}'_3$ together with the epipoles in $\mathbf{P}'_i$ and $\mathbf{P}'_j$. Let $\mathbf{e}'_i$ and $\mathbf{e}'_j$ be these epipoles. If $\mathbf{e}'_i$ and $\mathbf{e}'_j$ are known, then $\mathbf{H}_{ij}$ can be computed, and hence $\mathcal{F}_1$ induced by Δ_i and Δ_j can be determined using

$$\mathcal{F}_1 \sim [\mathbf{e}'_j]_\times \mathbf{H}_{ij}, \text{ or } \mathcal{F}_1 \sim \mathbf{H}_{ij}^{-T} [\mathbf{e}'_i]_\times. \tag{5}$$

Similarly, $\mathcal{F}_2$ induced by Δ_m^k and Δ_n^k is computed as

$$\mathcal{F}_2 \sim [\mathbf{e}'_n]_\times \mathbf{H}_{mn}, \text{ or } \mathcal{F}_2 \sim \mathbf{H}_{mn}^{-T} [\mathbf{e}'_m]_\times, \tag{6}$$

where $\mathbf{e}'_m$ and $\mathbf{e}'_n$ are the epipoles on virtual cameras $\mathbf{P}'_m$ and $\mathbf{P}'_n$, and $\mathbf{H}_{mn}$ is the induced homography.

The difficulty with (5) and (6) is that the epipoles $\mathbf{e}'_i$, $\mathbf{e}'_j$, $\mathbf{e}'_m$ and $\mathbf{e}'_n$ are unknown, and cannot be computed directly from the triplet correspondences. Fortunately, however, the epipoles can be closely approximated as described below.

Proposition 3 *If the exterior orientation of* $\mathbf{P}_1$ *is related to that of* $\mathbf{P}_2$ *by a translation, or by a rotation around an axis that lies on the axis planes of* $\mathbf{P}_1$*, then under the assumption:*

$$\mathbf{e}'_i = \mathbf{e}'_j = \mathbf{e}_1, \quad \mathbf{e}'_m = \mathbf{e}'_n = \mathbf{e}_2, \tag{7}$$

we have:

$$\mathcal{E}(\hat{\mathcal{F}}_1, \hat{\mathcal{F}}_2) = 0. \tag{8}$$

Under more general motion, the equalities in (7) become only approximate. However, we shall see in section 4.1.1 that this approximation is inconsequential in action recognition for a wide range of practical rotation angles. As described shortly, using equation (4) and the fundamental matrices $\mathcal{F}_1$ and $\mathcal{F}_2$ computed for every non-degenerate triplet, we can define a similarity measure for matching pose transitions $I_i \rightarrow I_j$ and $J_m^k \rightarrow J_n^k$.

Degenerate triplets: A homography cannot be computed from four correspondences if three points are collinear. Even when three image points are close to collinear the problem becomes ill-conditioned. We call such triplets as degenerate, and simply ignore them in matching pose transitions. This does not produce any difficulty in practice, since with 11 body point representation used in this paper (see Fig. 2), we obtain 165 possible triplets, the vast majority of which are in practice non-degenerate. A special case is when the epipole is close to or at infinity, for which all triplets would degenerate. We solve this problem by transforming the image points in projective space in a manner similar to Zhang et al. [19]. The idea is to find a pair of projective transformations $\mathbf{Q}$ and $\mathbf{Q}'$, such that after transformation the epipoles and transformed image points are not at infinity. Note that these transformations do not affect the projective equality in Proposition 2.

3.1.2 Algorithm for Matching Pose Transitions

The algorithm for matching two pose transitions $I_i \rightarrow I_j$ and $J_m^k \rightarrow J_n^k$ is as follows:

1. Compute $\mathbf{F}, \mathbf{e}_1, \mathbf{e}_2$ between image pair $\langle I_i, I_j \rangle$ and $\langle J_m^k, J_n^k \rangle$ using the method proposed in [7].
2. For each non-degenerate triplet Δ_ℓ that projects onto $\Delta_i, \Delta_j, \Delta_m^k$ and Δ_n^k in I_i, I_j, J_m^k and J_n^k, respectively, compute $\hat{\mathcal{F}}_1, \hat{\mathcal{F}}_2$ as described above from (5), (6) and (7), and compute $e_\ell = \mathcal{E}(\hat{\mathcal{F}}_1, \hat{\mathcal{F}}_2)$ from equation (4).
3. Compute the average error over all non-degenerate triplets using
$$E(I_i \rightarrow I_j, J_m^k \rightarrow J_n^k) = \frac{1}{L} \sum_{\ell=1...L} e_\ell, \tag{9}$$
where L is the total number of non-degenerate triplets.
4. If $E(I_i \rightarrow I_j, J_m^k \rightarrow J_n^k) < E_0$, where E_0 is some threshold, then the two pose transitions are matched. Otherwise, the two pose transitions are classified as mismatched.

3.2. Action Recognition

Given two sequences $A = \{I_{1...n}\}$ and $B = \{J_{1...m}\}$, we match or 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. This is solved by dynamic programming, which has been proven successful in sequence alignment, and its application in action recognition can also be found in [11]. The key is to define $S(.)$ based on matching pose transitions: $S(i,j) = \tau - E(I_{i\rightarrow r_1}, J_{j\rightarrow r_2})$, where $(r_1, r_2) = \underset{r_1, r_2}{\operatorname{argmin}}\{\underset{s_1, s_2}{\min} E(I_{r_1\rightarrow s_1}, J_{r_2\rightarrow s_2})\}$, $r_1, s_1 \in [1, n]$ and $r_2, s_2 \in [1, m]$. The matching score of sequences A and B is then defined by $\mathscr{S}(A, B) = \underset{\psi}{\max} \sum_{i=1}^n S(i, \psi(i))$.

To solve the action recognition problem, we need a reference sequence (a sequence of 2D poses) for each known action, and maintain an action database of K actions, $DB = \{J_t^1\}, \{J_t^2\}, \ldots, \{J_t^K\}$. To classify a given test sequence $\{I_t\}$, we match $\{I_t\}$ against each reference sequence in DB, and classify $\{I_t\}$ as the action of best-match, say $\{J_t^k\}$, if $\mathscr{S}(\{I_t\}, \{J_t^k\})$ is above a threshold T. Due to the use of view-invariant fundamental ratios, our solution is invariant to camera intrinsic parameters and viewpoint. To ensure the approximation of epipoles discussed above, reference sequences from 2-3 viewpoints may be used for each action.

4. Experimental Results and Discussion

We first examine our method on semi-synthetic data, and then test our solution on real video data.

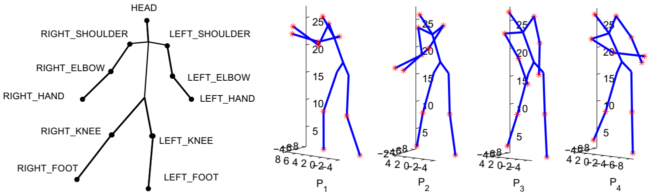


Figure 2. Left: Our body model. Right: Experiment on view-invariance. Two different pose transitions $P_1 \rightarrow P_2$ and $P_3 \rightarrow P_4$ from a golf swing action are used.

4.1. Analysis based on motion capture data

We generated our data based on the CMU Motion Capture Database, which consists of 3D motion data for a large number of human actions. We generated the semi-synthetic data by projecting 3D points onto images through synthesized cameras. In other words, our test data consist of video sequences of true persons, but the cameras are synthetic, resulting in semi-synthetic data to which various levels of noise were added. Instead of using all body points provided in CMU's database, we employed a body model that consists of only eleven points, including head, shoulders, elbows, hands, knees and feet (see Fig.2). This model is also used in the experiments in section 4.2.

4.1.1 Testing View Invariance

We selected four different poses P_1, P_2, P_3, P_4 from a golf swinging sequence (see Fig.2). We then generated two cameras as shown in Fig.3 (a): camera 1 was placed at an arbitrary viewpoint (marked by red color), with focal length $f_1 = 1000$; camera 2 was obtained by rotating camera 1 around an axis on x-z axis plane of camera 1 (colored as green), and a second axis on y-z axis plane of camera 1 (colored as blue), and changing focal length as $f_2 = 1200$. Let I_1 and I_2 be the images of poses P_1 and P_2 on camera 1 and I_3, I_4, I_5 and I_6 the images of poses P_1, P_2, P_3 and P_4 on camera 2, respectively. Two sets of pose similarity errors were computed at all camera positions shown in Fig.3 (a): $E(I_1 \rightarrow I_2, I_3 \rightarrow I_4)$ and $E(I_1 \rightarrow I_2, I_5 \rightarrow I_6)$. The results are plotted in Fig.3 (b) and (c), which show that, when two cameras are observing the same pose transitions, the error is zero regardless of their different viewpoints, confirming proposition 3.

Similarly, we fixed camera 1 and moved camera 2 on a sphere as shown in Fig.3 (d). The errors $E(I_1 \rightarrow I_2, I_3 \rightarrow I_4)$ and $E(I_1 \rightarrow I_2, I_5 \rightarrow I_6)$ are shown in Fig.3 (e) and (f). Under this more general camera motion, the pose similarity score of corresponding poses is not always zero, since the epipoles in equations (5) and (6) are approximated. However, this approximation is inconsequential in most situations, because the error surface of different pose transitions is in general above that of corresponding pose transitions. Fig.3 (h) shows the regions (black colored) where approximation is invalid. These regions correspond to the situation that the angles between camera orientations is around 90 degrees, which usually implies severe self-occlusion and lack of corresponding points in practice. The experiments on real data in section 4.2 also show the validity of this approximation under practical camera viewing angles.

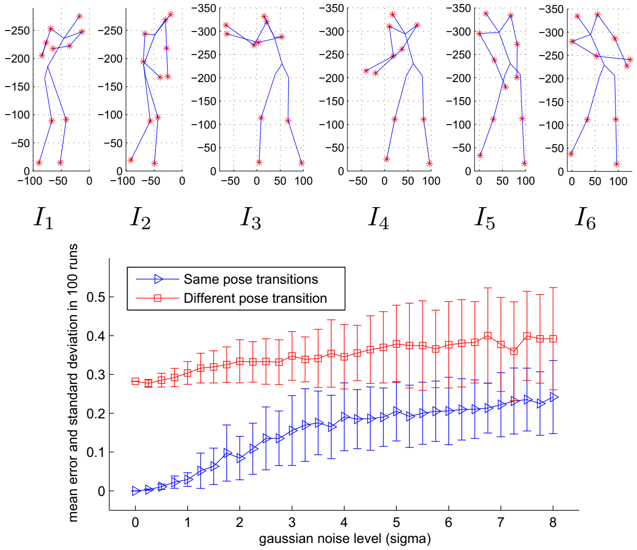


Figure 4. Robustness to noise: I_1 and I_2 are the images in camera 1, and I_3, I_4, I_5 and I_6 are the images in camera 2. Same and different actions are distinguished unambiguously for $\sigma < 4$

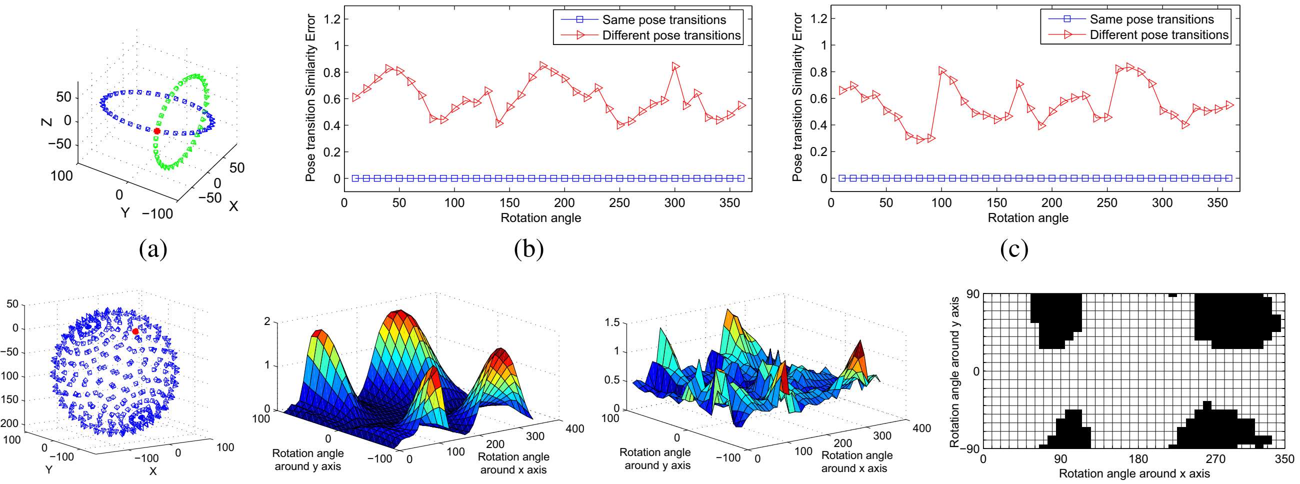
4.1.2 Testing Robustness to Noise

Without loss of generality, we used the four poses in Fig.2 to analyze the robustness of our method to noise. Two cameras with different focal lengths and viewpoints were examined. As shown in Fig.4, I_1 and I_2 are the images of poses P_1 and P_2 on camera 1 and I_3, I_4, I_5 and I_6 are the images of P_1,P_2,P_3 and P_4 on camera 2. We then added Gaussian noise to the image points, with σ increasing from 0 to 8. The errors $E(I_1 \rightarrow I_2, I_3 \rightarrow I_4)$ and $E(I_1 \rightarrow I_2, I_5 \rightarrow I_6)$ were computed. For each noise level, the experiment was repeated for 100 independent trials, and the mean and standard deviation of both errors were calculated (see Fig.4). As shown in the results, the two cases are distinguished unambiguously until σ increases to 4.0, i.e., up to possibly 12 pixels. Note that the image sizes of the subject were about 200×300, which implies that our method performs remarkably well under high noise.

4.1.3 Performance in Action Recognition

We selected 5 classes of actions from CMU's MoCap dataset: walk, jump, golf swing, run, and climb. Each action class is performed by 3 actors, and each instance of 3D action is observed by 17 cameras, as shown in Fig.5. The focal lengths were changed randomly in the range of 1000 ± 300. Fig.6 shows an example of a 3D pose observed from 17 viewpoints.

Our dataset consists of totally 255 video sequences, from which we generated a reference action Database (DB) of 5



(a) (b) (c)

(d) (e) (f) (g)

Figure 3. Analysis of view invariance. (a) Camera 1 is marked in red, and all positions of camera 2 are marked in blue and green. (b) Errors for same and different pose transitions when camera 2 is located at viewpoints colored as green in (a). (c) Errors of same and different pose transitions when camera 2 is located at viewpoints colored as blue in (a). (d) General camera motion: Camera 1 is marked as red, and camera 2 is distributed on a sphere. (e) Error surface of same pose transitions for all distributions of camera 2 in (d). (f) Error surface of different pose transitions for all distribution of camera 2 in (d). (g) The regions of confusion for (d) marked in black (see text).

	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	1	1	0	1	1	1	2	0	1	13	2	0	4	11	7	1
Accuracy (%)	100	93.3	93.3	100	93.3	93.3	93.3	86.7	100	93.3	13.3	86.7	100	73.3	26.7	53.3	93.3

Table 1. Recognition accuracy for various viewpoints illustrated in Fig. 6.

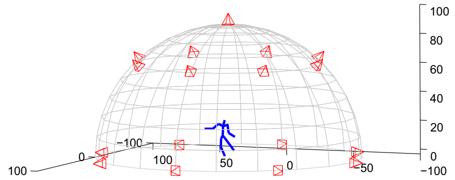


Figure 5. The distribution of cameras used to evaluate view-invariance and camera parameter changes.

video sequences, i.e. one video sequence for each action class . The rest of the dataset was used as test data, and each sequence was matched against all actions in the DB and classified as the one with highest score. For each sequence matching, 10 random initializations were tested and the best score was used. The overall classification accuracy for all viewpoints is 81.60%, with very low accuracy at viewpoints 11, 14, 15, 16, which correspond to severe viewing angles from below or above the actor. This is consistent with observations pointed out in section 4.1.1. Excluding these viewpoints, the classification accuracy increases to 94.21%.

4.2. Results on real data

We collected video data from Internet, consisting of 56 sequences of 8 classes of actions. Fig.7 (a) shows an example of matching action sequences. The frame rates and viewpoints of two sequences are different, and two players perform golf-swing action at different speeds. The accumulated score matrix and back-tracked path in dynamic programming are shown in Fig.7 (c). Another result on tennis-serve sequences is shown in Fig.7 (b) and (d). More details and more results are included in the supplementary video.

Ground-true actions	Recognized as action							
	#1	#2	#3	#4	#5	#6	#7	#8
#1	3							
#2	1	10						
#3			5					
#4				7				
#5					3			
#6					1	6		
#7							3	
#8								9

Table 2. Confusion matrix. The actions are denoted by numbers: 1 - ballet fouette, 2 - ballet spin, 3 - pushup, 4 - golf swing, 5 - one handed tennis backhand, 6 - two handed tennis backhand, 7 - tennis forehand, 8 - tennis serve. The diagonal nature of the matrix indicates high accuracy.

We built an action database DB by selecting one sequence for each action; the rest were used as test data, and were matched against all actions in the DB. An action was recognized as the one with highest matching score. The confusion matrix is shown in Table 2, which indicates an overall 95.83% classification accuracy for real data.

5. Conclusion

There are three major contributions in this paper: (1) we introduce the concept of *fundamental ratios* and apply it to

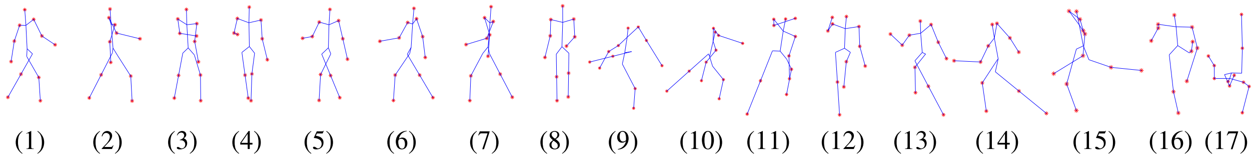
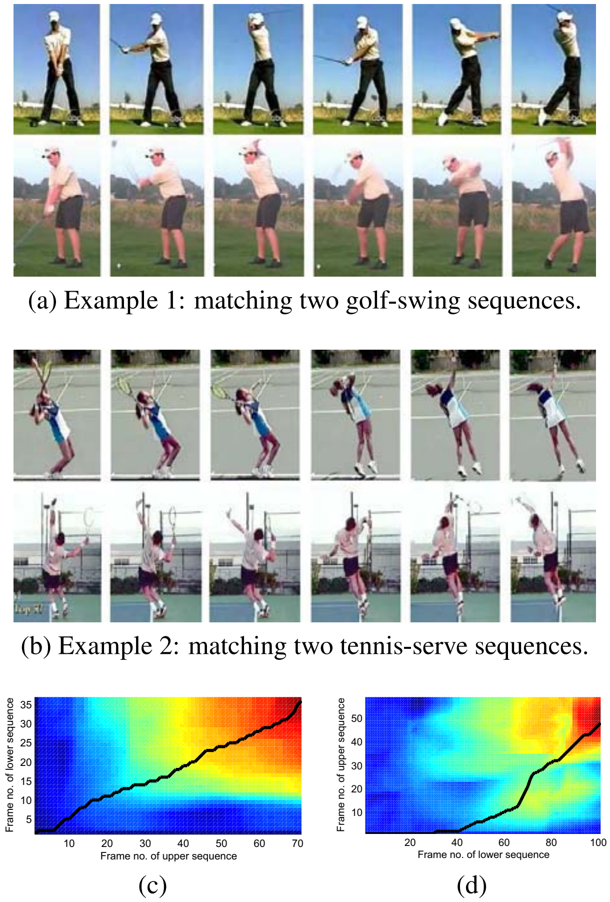


Figure 6. A pose observed from 17 viewpoints. Note that only 11 body points in red color are used. The stick shapes are shown here for better illustration of pose configuration and extreme variability being handled by our method.


(a) Example 1: matching two golf-swing sequences.

(b) Example 2: matching two tennis-serve sequences.

Figure 7. Examples of matching action sequences. (a) and (b) are two examples in golf-swing and tennis-serve actions. (c) and (d) show the accumulated score matrices and backtracked paths, resulting in the alignments shown in (a) and (b), respectively.

action recognition; (2) we propose to compare transitions of two poses, which encodes temporal information of human motion and keeps the problem at its atomic level; (3) we propose to break a human pose into a set of triplets and represent a human action by the motion of planes of triplets. This converts the study of non-rigid human motion into that of multiple rigid motions of planes, making it thus possible to apply well-studied rigid motion concepts, and providing a novel direction to study articulated motion.

Acknowledgement: We would like to thank Dr. Xiaochun Cao and Dr. Imran Junejo for their insightful discussions and ideas.

References

[1] M. Ahmad and S. Lee. HMM-based Human Action Recognition Using Multiview Image Sequences. *Proc. ICPR*, pages 263–266, 2006. 1

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

[3] X. Cao, J. Xiao, and H. Foroosh. Camera Motion Quantification and Alignment. *Proc. ICPR*, 13–16, 2006. 2

[4] F. Cuzzolin. Using Bilinear Models for View-invariant Action and Identity Recognition. *CVPR*, 1701–1708, 2006. 1

[5] D. Gavrila. Visual analysis of human movement: A survey. *CVIU*, 73(1):82–98, 1999. 1

[6] A. Gritai, et al. On the use of anthropometry in the invariant analysis of human actions. *Proc. ICPR*, 2, 2004. 1, 2

[7] R. Hartley. In defence of the 8-point algorithm. *Proc. ICCV*, 00:1064, 1995. 3

[8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 1, 2

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

[10] T. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 104(2-3):90–126, 2006. 1

[11] V. Parameswaran and R. Chellappa. View Invariance for Human Action Recognition. *IJCV*, 66(1):83–101, 2006. 1, 3

[12] C. Rao, et al. View-Invariant Representation and Recognition of Actions. *IJCV*, 50(2):203–226, 2002. 1

[13] Y. Sheikh et al. Exploring the Space of a Human Action. *Proc. ICCV*, 1, 2005. 1

[14] T. Syeda-Mahmood, A. Vasilescu, S. Sethi, I. Center, and C. San Jose. Recognizing action events from multiple viewpoints. *Proc. IEEE Workshop DREV*, pages 64–72, 2001. 1

[15] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36:585–601, 2003. 1

[16] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *CVIU*, 104(2-3):249–257, 2006. 1

[17] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. *Proc. CVPR*, 1, 2005. 1, 2

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