Action Recognition based on Homography Constraints

Yuping Shen, Nazim Ashraf and Hassan Foroosh
Computational Imaging Lab., University of Central Florida, Orlando, FL 32816
{*ypshen,nazim,foroosh*}*@eecs.ucf.edu*

Abstract

In this paper, we present a new approach for view-invariant action recognition using constraints derived from the eigenvalues of planar homographies associated with triplets of body points. Unlike existing methods that study an action as a whole, or break it down into individual poses, we represent an action as a sequence of pose transitions*. 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 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. 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 camera parameters.*

1. Introduction

Human action recognition has been the subject of extensive studies in the past, highlighted in recent survey papers such as [4, 8, 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. 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 coplanarity of a subset of body points [9], etc.

There are mainly two lines of research to tackle view-invariance: One is based on using multiple cameras, such as [16, 7, 2], 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 [10], or projective [9], 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, 5].

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. We represent an action as a set of non-rigid *pose transitions* defined by triplets of points - that is we break down further each pose into a set of point-triplets and find invariants for the motion of these triplets across two frames.

2. Action Recognition Based on Homography Constraints

2.1. Representation of Human Action

In the study of human motion, set of body points is widely used to represent a human body pose. Other representations of pose include silhouette [1], optical flow [3] and local space-time features [12]. Based on their different use of pose representation, existing research works regard an human action either as a whole, or as a sequence of individual poses.

In this paper, we represent a human body pose $\mathcal{P}$ by M body points: $\mathcal{P} = \{\mathbf{m}_{i=1...M}\}$. These points, which are the only inputs to our algorithm, can be obtained by using articulated object tracking techniques such as [11]. Further discussions on articulated object tracking can be found in [4, 8], 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.

Suppose an action sequence $\mathbf{A}$ consists of T frames: $\{\mathcal{P}_1^A, \ldots, \mathcal{P}_T^A\}$. With an arbitrarily pose $\mathcal{P}_r^A$ selected as reference, $\mathbf{A}$ is decomposed into a sequence of pose transitions: $\mathbf{A}(r) = \{\mathcal{P}_i^A \rightarrow \mathcal{P}_r^A | i = 1..M, i \neq r\}$. With this representation, comparison of two action sequences reduces to examining the similarities of pose transitions, as described in the following sections.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

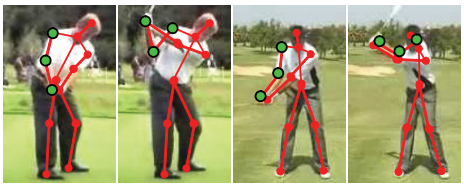


Figure 1. An example of similar pose transitions.

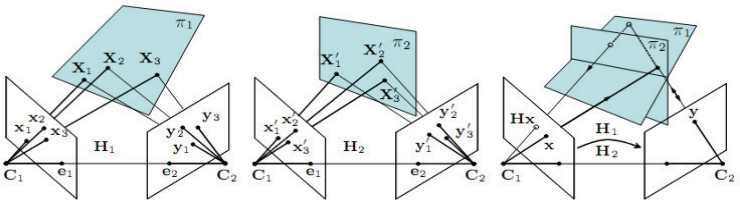


Figure 2. Homographies induced by a moving triplet of points.

2.2. Matching Pose Transitions

Supposed we are given two pose transitions $I_{1\to 2}$[1] and $J_{i\to j}$ (see Fig. 1 for example). 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.

Consider the case that $I_{1\to 2}$ corresponds to $J_{i\to j}$. $I_{1,2}$ 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 computation of $\mathbf{F}$ has been well studied in decades, e.g, [6]. Note that $\mathbf{F}$ does not correlate the entire scene, but only the body points of the subjects.

2.2.1. Homographies Induced by a Body-Points Triplet. Let us now consider an arbitrary triplet of 3D body points (see highlighted points in Fig. 1 for example), $\mathbf{\Delta} = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$, which corresponds to $\Delta_1 = \langle \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \rangle$ in I_1 and $\Delta_i = \langle \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3 \rangle$ in J_i. After the pose transformation, $\mathbf{\Delta}$ transforms to $\mathbf{\Delta}' = \{\mathbf{X}'_1, \mathbf{X}'_2, \mathbf{X}'_3\}$, which corresponds to $\Delta_2 = \langle \mathbf{x}'_1, \mathbf{x}'_2, \mathbf{x}'_3 \rangle$, in I_2 and $\Delta_j = \langle \mathbf{y}'_1, \mathbf{y}'_2, \mathbf{y}'_3 \rangle$ in J_j, as illustrated in Fig. 2. $\mathbf{\Delta}$ 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 (Note that the total number of available triplets is $\binom{n}{3}$ for n body points). A special case is when the epipole is at or close to infinity, all triplets then may be regarded as 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, which is similar to [18]. The idea is to find the projective transformation $\mathbf{P}$ and $\mathbf{P}'$ for each image, such that after transformation the epipoles and image points are finite.

2.2.2. Constraints on Homographies due to Moving Triplets. As described above, during a pose transition, the motion of a triplet $\mathbf{\Delta} \to \mathbf{\Delta}'$ 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

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

As shown in Fig. 2, $\mathbf{H}$ first maps a point $\mathbf{x}$ on I_1 (or I_2) to $\mathbf{y}$ on J_i (or J_j) through π_1, and then transforms it back to I_1 (or I_2) as $\mathbf{Hx}$ through π_2. It can be readily verified either algebraically or from Fig. 2 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 corresponding triplets of human body points in the two video sequences undergo the same motion, then the homography* $\mathbf{H}$ *reduces to a planar homology, and hence two of its eigenvalues must be equal.*

One can easily verify that, when two corresponding triplets undergo different motion, the homographies $\mathbf{H}_1$ and $\mathbf{H}_2$ computed as above are not compatible with the ground-truth fundamental matrix between two views, hence $\mathbf{H}$ is not a planar homology. Therefore, 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, 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, this constraint of equality of two eigenvalues for the same pose transition can be expressed by defining the following error function on $\mathbf{H}$:

[1] For brevity of notation, we denote transition $I_1 \to I_2$ as $I_{1\to 2}$.

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

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 triplets of non-collinear points provides a measure of distance between $I_{1\to 2}$ and $J_{i\to j}$:

$$\mathcal{E}(I_{1\to 2}, J_{i\to j}) = 1 - \underset{\text{all } \Delta_i}{\text{Median}}(E(\mathbf{H})). \quad (2)$$

$\mathcal{E}(I_{1\to 2}, J_{i\to j})$ is maximal for similar pose transitions, and is invariant to camera calibration matrix and viewpoint variations.

2.3. Action Recognition

Given two sequences $A = \{I_{1\ldots n}\}$ and $B = \{J_{1\ldots m}\}$, we match or align A and B by seeking the optimal mapping $\psi : A \to B$ 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 [9]. The key is to define $S(.)$ based on matching pose transitions: $S(i,j) = \tau - \mathcal{E}(I_{i\to r_1}, J_{j\to r_2})$, where τ is a constant threshold, $(r_1, r_2) = \underset{r_1,r_2}{\text{argmin}}\{\underset{s_1,s_2}{\min} E(I_{r_1\to s_1}, J_{r_2\to s_2})\}$, $r_1, s_1 \in [1,n]$ and $r_2, s_2 \in [1,m]$. The matching score of A and B is then defined by $\mathscr{S}(A,B) = \underset{\psi}{\max} \sum_{i=1}^{n} S(i,\psi(i))$. In other words, a pair of reference poses $\langle r_1, r_2 \rangle$ are found first, then two pose-transition sequences $A(r_1)$ and $B(r_2)$ are aligned using DP.

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 distant in equation 2, our solution is invariant to camera intrinsic parameters and viewpoint.

3. Experimental Results

In this section, we validate our approach on both semi-synthetic and real data.

3.1. Results 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. Instead of using all body points provided in CMU's database, we employed a body model that consists of only 11 joints and end points, including head, shoulders, elbows, hands, knees and feet.

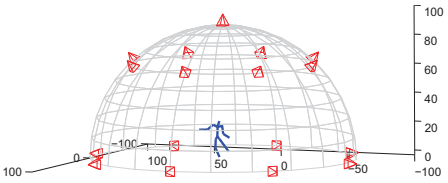


Figure 3. The distribution of cameras used in semi-synthetic data.

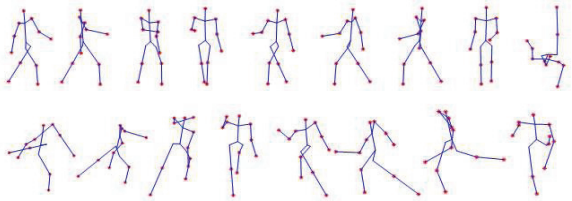


Figure 4. A pose observed from 17 viewpoints.

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.3. The focal lengths were changed randomly in the range of 1000 ± 300. Fig.4 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 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 classification results are shown in Table 3.1. The number in row 1, column 5 means that two of walking sequences are misclassified as climbing. The overall classification accuracy is about 92%.

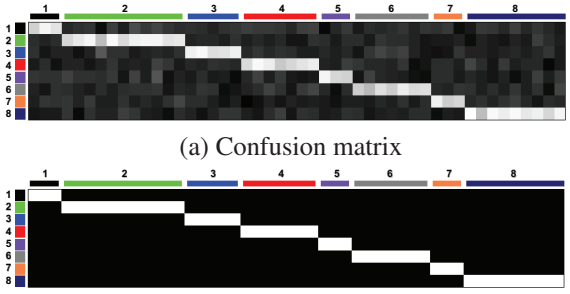
3.2. Results on Real Data

We also evaluated our method on a dataset of real video sequences. To best simulate the situations in real life, we collected these videos from Internet, coming from a variety of sources. The collected dataset consists of 56 sequences of 8 actions[2]: 4 of ballet fouettes, 12 of ballet spin, 6 of push-up, 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

[2] see http://cil.cs.ucf.edu/ar.html for details.

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 1. Confusion matrix: Large values on the diagonal entries indicate accuracy.



(a) Confusion matrix

(b) Recognition result

Figure 5. Action recognition results based on the confusion matrix (see text for details). All actions are correctly classified.

subjects, observed from different unknown viewpoints. We used the same human model as in section 3.1.

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. 5 (a), where light colors represent similar actions and dark colors represent dissimilar actions. 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. The recognition result is shown in Fig. 5 (b), where the black block in each column indicates the recognized action for each test sequence. As shown in the results, our method provides a successful recognition of various actions by different subjects, regardless of large differences in unknown camera intrinsic parameters and viewpoints. Our method also accommodates substantial self-occlusions and minor camera motions.

4. Conclusion and Discussion

In summary, we present a new approach for view-invariant action recognition using constraints derived from the eigenvalues of planar homologies associated with triplets of body points. We treat an action as a sequence of pose transitions, and compare two actions by matching their pose transitions - instead of pose-to-pose comparison, we compare the pose transitions between two sequences. For two similar *pose transitions*, each triplet of non-collinear body points induce two homographies across time, which should satisfy the constraint of equality of eigenvalues. This constraint is independent of camera calibration matrix and viewpoint, thus provides a cue for view-invariant action recognition. The proposed method is evaluated on both semi-synthetic and real data.

References

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

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

[3] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *ICCV*, pages 726–733, 2003.

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

[5] A. Gritai, Y. Sheikh, and M. Shah. On the use of anthropometry in the invariant analysis of human actions. *ICPR*, 2, 2004.

[6] Q.-T. Luong and O. D. Faugeras. The fundamental matrix: theory, algorithms, and stability analysis. *International Journal of Computer Vision*, 17, 1996.

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

[8] 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.

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

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

[11] J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, pages 612–617, 1995.

[12] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. *ICPR*, 3, 2004.

[13] Y. Sheikh and M. Shah. Exploring the Space of a Human Action. *ICCV*, 1, 2005.

[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.

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

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

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

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