# Recognizing Human Actions in Videos Acquired by Uncalibrated Moving Cameras

Alper Yilmaz
School of Computer Science
University of Central Florida
yilmaz@cs.ucf.edu

Mubarak Shah
School of Computer Science
University of Central Florida
shah@cs.ucf.edu

## Abstract

*Most work in action recognition deals with sequences acquired by* stationary *cameras with* fixed viewpoints*. Due to the camera motion, the trajectories of the body parts contain not only the motion of the performing actor but also the motion of the camera. In addition to the camera motion, different viewpoints of the same action in different environments result in different trajectories, which can not be matched using standard approaches. In order to handle these problems, we propose to use the multi-view geometry between two actions. However, well known epipolar geometry of the static scenes where the cameras are stationary is not suitable for our task. Thus, we propose to extend the standard epipolar geometry to the geometry of dynamic scenes where the cameras are moving. We demonstrate the versatility of the proposed geometric approach for recognition of actions in a number of challenging sequences.*

## 1. Introduction

During the last two decades, a large number of research articles have been published on the recognition of human actions. This popularity is mainly due to the occurrence of actions in many real world applications such as surveillance, video classification and content based retrieval. Regardless, both of these tasks still remain outstanding challenges in the Vision Community.

A common approach taken by the researchers is to perform action recognition in 2-D, such as, use of motion trajectories [16], optical flow vectors [4] and silhouettes [3]. For instance, for recognizing facial expressions, Black and Yacoob [2] computed the affine motion of the bounding boxes around the eyes, the eyebrows and the mouth. The variations in the affine parameters are shown to capture facial changes during an expression. Yang et al. [21] also used the affine transformation computed between the corresponding segments in consecutive frames for sign language recognition. Before their work the same problem was addressed by Starner and Pentland [17], where the authors used the bounding boxes around the hands to training HMMs which model the states of the hand during the action. Efros et al. [4] used the optical flow computed in the bounding boxes of the objects to represent the actions. Similarly, Polana and Nelson [14] generated the statistics of the normal flow from the spatio-temporal cube to represent the motion content during an action. Laptev and Lindeberg [12] used temporal and spatial image gradients to find descriptors of an action. Instead of using trajectory or bounding boxes, Bobick and Davis [3] used object silhouettes to model the action. A stack of such silhouettes, which provides a motion history, was called the a temporal template. Note that all of the aforementioned work can not recognize two different views of same action.

The viewpoint of the camera used to acquire the execution of the action plays an important role. This is mainly due to the fact that, appearance of the same action may drastically vary from one viewpoint to the other. Thus, use of the *stationary cameras with fixed viewpoints* has become a standard. In the case, when the cameras are not fixed and are moving independently, the variation in the appearance of an action is even more drastic. For instance, when the cameras move, the camera motion induces false motion in the motion trajectories of the actor. In fact, to the best of the authors' knowledge there is no work in action recognition employing sequences acquired by moving cameras.

Recently, view invariance issue, which relaxes fixed viewpoint constraint but relies on stationary cameras, has become an active topic of research in action recognition. In [16], authors represented the actions by dynamic instants which are computed from the curvature maxima of the motion trajectories. In similar vein, authors of [7] used thirteen trajectories of landmark points on the human body. Yilmaz and Shah [22] represented the action by a set of descriptor computed from a spatio-temporal action volume created from a set of object contours. To perform view invariant

recognition of actions, all three approaches [16, 7, 22] use the epipolar geometry between the views of two stationary cameras. In [13], Parameswaran et al. used five landmark points on the human body which are conjectured to form a plane in real world during execution of the action. From this plane, they compute a set of projective invariants to match different poses of an actor from different viewpoints.

In this paper we address the following question: *How can we recognize actions when the cameras are moving?* It is obvious that the standard epipolar geometry can not be used for moving cameras [20]. An unattractive solution to recognize actions in this scenario is to recover the epipolar geometry between the corresponding frames in two views independently[1]. However, this is not attractive for a number of reasons: 1) High computational cost (number of parameters to be estimated increase linearly as the number of frames increase), 2) Fundamental matrices in consecutive frames have to be temporally related to each other. We will call the relation between consecutive fundamental matrices: temporal consistency.

In [1], Avidan and Shashua use the trifocal tensor to guarantee temporal consistency between two consecutive static fundamental matrices. Recently, research dealing with dynamic scenes (both the scene and the cameras are moving [6]) has become more active. For instance, in case of recovering the shape from motion, Wolf and Shashua [20] constrain the motion of the objects, such that the objects follow straight paths with "constant speed" or "constant acceleration". Under these constraints, they showed that using velocities as additional dimensions to the spatial space reduces the geometry of a dynamic scene to the well-known epipolar geometry. In the case of actions, straight path, constant speed or constant acceleration constraints are not satisfied. In order to relax these constraints, in this paper, we propose to model the variation in the epipolar geometry of dynamic scenes by means of a temporal fundamental matrix (TFM) which is $3 \times 3$ matrix function, $\mathcal{F}(t)$, where $t$ denotes time. TFM is derived by analyzing the effect of the camera motion (rotational and translational motion in 3D) on the scene geometry. Using TFM, we formulate the action recognition problem in terms of the quality of the recovered scene geometry. Given the trajectories of the landmarks on the human body, this is performed in two steps:

1. Labeled trajectories are organized to form a linear system of equations $\mathbf{M}\mathbf{f} = 0$ to estimate unknowns $\mathbf{f}$,

2. Action recognition is performed by simultaneously minimizing the geometric error of the recovered geometry and maximizing the quality of the linear system[2].

[1]Note that, independent camera motion is required to observe independent fundamental matrices at each time instant. For instance, for a pair of moving cameras on a stereo rig, where cameras are fixed with respect to each other, the fundamental matrix has constant components. This is due to the zero relative motion between the camera pair.

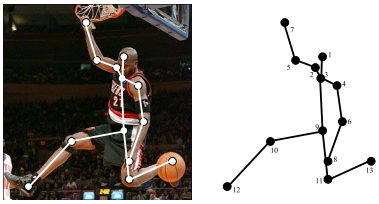


**Figure 1. Point based representation of the actor's posture. Johansson [11] has shown that point based representation provide sufficient information to infer the action performed by the actor.**

Demonstrated under a variety of challenging experiments, proposed action recognition is shown to be view invariant, robust to camera motion, and least affected by different execution styles of actions.

The paper is organized as follows. In Sec. 2, we discuss the action representation used in the paper. Section 3 discusses the proposed action matching approach. In Sec. 4, we demonstrate the versatility of our approach for two different applications. Finally, we conclude in Sec. 5.

## 2. Representation of Human Actions

A complete action representation might be the set of all three-dimensional points on a performing actor. During the execution of the action, these three-dimensional points generate four-dimensional trajectories $\Gamma_{4D}$ in space and time $(X, Y, Z, t)$, which can be projected to three-dimensional trajectories $\Gamma_{3D}$ in the spatio-temporal space $(x, y, t)$ by

$$\Gamma_{3D} = \underbrace{\begin{bmatrix} a & b & c & 0 & d \\ e & f & h & 0 & h \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}}_{\Pi} \Gamma_{4D}, \qquad (1)$$

where both $\Gamma_{3D}$ and $\Gamma_{4D}$ are in homogenous coordinates, and $\Pi$ is a $4 \times 5$ matrix. Having space-time trajectories of all the points is not practical. The question at hand is: "Is there a subset of points on the human body that is adequate for perception of an action?" Luckily, perception of human actions has been well-studied in psychology. In an experiment by Johansson [11], it was shown that a set of bright spots attached to the joints of an actor dressed in black provide sufficient information to infer the action being performed in front of a dark background. Note that, the collection of

[2]The quality of a linear system is computed from the condition number.

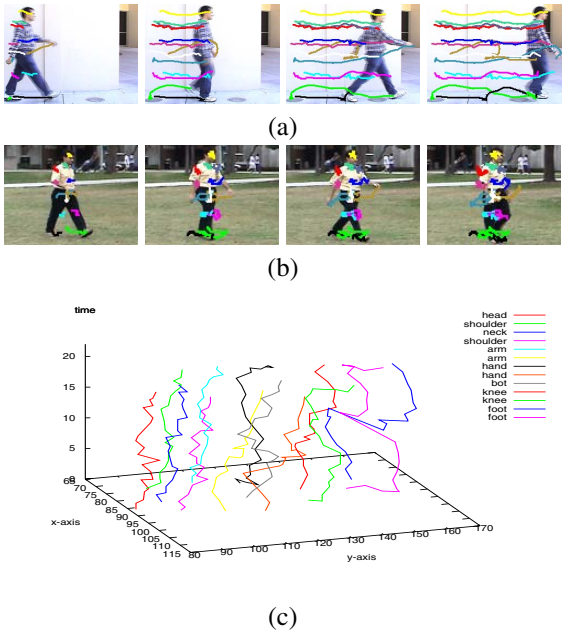



**Figure 2. Trajectories of the walking actions captured using (a) a stationary camera, and (b) a moving camera. (c) Trajectories of 13 landmark points of the walking sequence in (b) in the spatio-temporal space.**

bright spots carry only the spatial information without any structural information. Relying on this result, we represent the posture of an action by thirteen points (landmark points) positioned on the joints of the actor (see Fig. 1 for the point based representation used in this discussion).

During the performance of an action, the $i^{th}$ landmark point generates the spatio-temporal trajectory $\Gamma^i_{3D} = (\mathbf{x}_1^\top, \mathbf{x}_2^\top \ldots \mathbf{x}_n^\top)$, where $n$ is the duration of the action and $\Gamma^i_{3D}$ is $3n \times 1$ vector. For clarity, we will remove subscript $3D$ in the following discussion. Each action $\mathbf{U}$ is represented by a collection of thirteen trajectories:

$$\mathbf{U} = (\Gamma_1^\top, \Gamma_2^\top \ldots \Gamma_{13}^\top). \quad (2)$$

We will call $\mathbf{U}$ matrix, the "action matrix".

In Fig. 2, we show two sequences of the walking action with the trajectories superimposed. The first sequence is captured using a stationary camera (part (a)) and second sequence is captured using a moving camera (part (b)). Due to the motion of the camera, it is evident that the trajectories in part (a) and part (b) do not appear similar. In Fig. 2c, we plot the trajectories of all the landmark points in part (b) in the spatio-temporal space (in particular the action representation given in equation (2)).

# 3. Matching Actions

A common approach among researchers who proposed view invariant action recognition is to adapt the machinery of multi-view geometry which is generally used in context of stereo [10] and structure from motion [23]. In this setting, action matching relies on the fact that two views of the same action performed by different actors results in similar trajectories obtained from a set of labeled points. Despite the success stories reported in the research papers, an important limitation of this well-known geometry is the requirement of stationary cameras, which is usually not satisfied in real world applications, such as retrieving an action from a movie or broadcast news video. A trivial solution to overcome the problems related to the camera motion is to compensate frame to frame global motion, such that resulting trajectories do not contain camera motion [15]. This however is not attractive for a number of reasons:

1. Compensating global motion has high computational cost.
2. Global motion compensation methods assume planar scenes and are generally suitable for distant views. However, actions are usually captured as closeup views and due to the motion parallax a poor motion estimation is obtained.
3. Compensating motion distorts the multi-view geometry properties by introducing artificial deformations.

An alternative approach is to recover the scene geometry by estimating the fundamental matrix independently for every frame[3]. Similar to the global motion estimation, this approach has high computational complexity. Moreover, additional constraints, such as the use of the trifocal tensors [1] is required to guarantee temporal consistency between the consecutive fundamental matrices.

In this section, we propose a novel action matching method based on the geometry of the dynamic scenes. Proposed method leverages the state of the art in action recognition by relaxing the stationary camera constraint.

## 3.1. Multi-View Geometry

In order to understand the geometry of a dynamic scene, we first start the discussion with the geometry of a static scene which has been well studied.

### 3.1.1 The geometry of a static scene

Static scene geometry captured from two stationary cameras is given in Fig. 3a. We will call this the "static epipolar

[3]For $n$ frames estimation of the scene geometry requires computation of $9n$ unknowns.

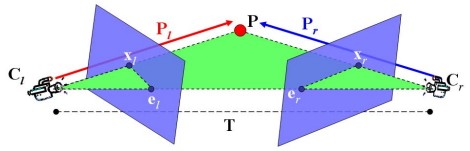



**Figure 3. Epipolar geometry of a static scene with two fixed cameras. $\mathrm{P}$ denotes a 3D point (a landmark point on the actor performing the action, $\mathrm{C}_l$ and $\mathrm{C}_r$ are left and right camera centers, $\mathrm{e}_l$ and $\mathrm{e}_r$ are the epipoles of the left and right image planes.**

geometry". In static epipolar geometry, the projection of a world point $\mathbf{P}$ to the left camera reference frame, $\mathbf{P}_l$, and right camera reference frame, $\mathbf{P}_r$ are related by [19]:

$$\mathbf{P}_r = \mathbf{R}\left(\mathbf{P}_l - \mathbf{T}\right), \tag{3}$$

where

$$\mathbf{R} = \mathbf{R}_r\mathbf{R}_l^\top \tag{4}$$

is the rotation from the left to the right camera reference frame, and the vector $\mathbf{T}$ defined below connects the camera centers:

$$\mathbf{T} = (T_x, T_y, T_z)^\top = \mathbf{T}_l - \mathbf{R}^\top\mathbf{T}_r. \tag{5}$$

An immediate result of equation (3) and the coplanarity constraint on the epipolar plane is the essential matrix $\mathcal{E}$ which satisfies:

$$\mathbf{P}_r^\top\left(\mathbf{RS}\right)\mathbf{P}_l = \mathbf{P}_r^\top\mathcal{E}\mathbf{P}_l = 0, \tag{6}$$

where $\mathbf{S}$ is a rank deficient matrix obtained from $\mathbf{T}$. The essential matrix in equation (6) can be extended to relate the image planes of the left and the right cameras by introducing the intrinsic camera parameters, $M_l$ and $M_R$, such that $\mathbf{x} = M_l\mathbf{P}_l$ and $\mathbf{x}' = M_r\mathbf{P}_r$, where $\mathbf{x}' = (x', y', 1)$, $\mathbf{x} = (x, y, 1)$ respectively are the homogeneous image coordinates in the left and the right views of the scene. Putting these relations in equation (6) we get:

$$\mathbf{x}'^T(M_r^{-T}\mathcal{E}M_l^{-1})\mathbf{x} = \mathbf{x}'^T\mathcal{F}\mathbf{x} = 0, \tag{7}$$

where $\mathcal{F}$ is called the static fundamental matrix [5].

### 3.1.2 The geometry of a dynamic scene

Dynamic scene geometry is not as simple as the static scene (see Fig. 4b). The geometry at each time instant may be different from the geometry at the previous time instant[4].

[4]At this point, we should note that observing different epipolar geometry is not dependent on the scene content (motion of the actor) [9].

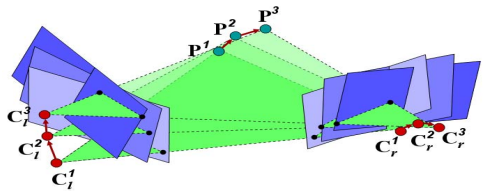



**Figure 4. Temporal epipolar geometry between two views for moving cameras and a moving landmark point on the actor. At each time instant the geometry changes, such that new epipoles and epipolar lines are observed. The superscripts denote the time and subscripts denote the left or the right camera. The arrows denote the direction of motion.**

To analyze the variation of this geometry, first, we introduce the motion of the camera in the 3D space. Let $\big(\Omega(0), \Theta(0)\big)$ denote the initial camera pose (rotation and translation), and let the camera move with rotational velocities $\omega_x(t)$, $\omega_y(t)$ and $\omega_z(t)$, and translational velocities $\theta_x(t)$, $\theta_y(t)$ and $\theta_z(t)$. Under the constraint that the camera motion is small[5], using Euler angles, one can show that the rotation matrix of the camera at time $t$ becomes:

$$\Omega(t) = \begin{bmatrix} 1 & -\sum_1^t \omega_z(t) & -\sum_1^t \omega_y(t) \\ \sum_1^t \omega_z(t) & 1 & -\sum_1^t \omega_x(t) \\ \sum_1^t \omega_y(t) & \sum_1^t \omega_x(t) & 1 \end{bmatrix} \Omega(0). \tag{8}$$

Similarly, it is easy to show that the translation vector of the camera at time $t$ is:

$$\Theta(t) = \begin{pmatrix} \sum_{i=0}^t v_x(t) & \sum_{i=0}^t v_y(t) & \sum_{i=0}^t v_z(t) \end{pmatrix}. \tag{9}$$

In the dynamic scene setting, both cameras move independently, such that $\Delta\Omega_l(1..t) \neq \Delta\Omega_r(1..t)$ and $\Delta\Theta_l(1..t) \neq \Delta\Theta_r(1..t)$, where subscripts denote left and right cameras. Thus, rotation from the left to the right camera reference frame given in equation (4) becomes:

$$\mathbf{R}(t) = \Omega_r(t)\Omega_l^\top(t). \tag{10}$$

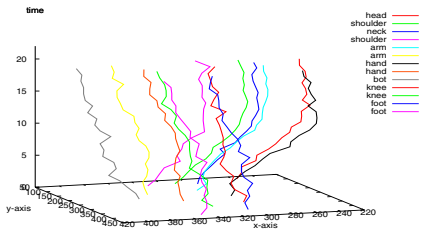
Similarly, equation (5) becomes:

$$\mathbf{T}(t) = \Theta_(t) - \mathbf{R}(t)\Theta_r(t). \tag{11}$$

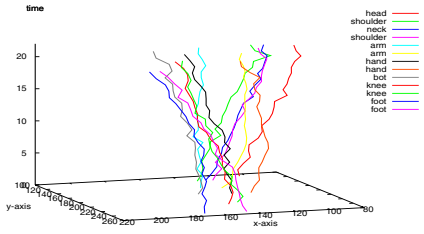
Based on these results the essential matrix of the dynamic scene is given by: $\tilde{\mathcal{E}}(t) = \mathbf{R}(t)\mathbf{S}(t)$, where $\mathbf{S}(t)$ is similar in construction to the $\mathbf{S}$ given for the static camera setting. Including the intrinsic camera parameters of the left and the right cameras fundamental matrix for the dynamic scene is:

$$\tilde{\mathcal{F}}(t) = M_r^{-\top}\left(\mathbf{R}(t)\mathbf{S}(t)\right)M_l^{-1}, \tag{12}$$

[5]$\omega_x(t) \approx 0$, $\omega_y(t) \approx 0$, $\omega_z(t) \approx 0$, $\theta_x(t) \approx 0$, $\theta_y(t) \approx 0$ and $\theta_z(t) \approx 0$




(a)




(b)

**Figure 5. Trajectories of the landmark points of two actors performing the picking up action which is captured from different viewpoints. In both sequences cameras undergo different motions. As seen the trajectories are not similar. (a) Camera is translating in the $x$ axis, (b) camera is rotating around the $z$ axis.**

where $M_l$ and $M_r$ are the intrinsic camera parameters[6]. We call $\tilde{\mathcal{F}}(t)$ "the temporal fundamental matrix" (TFM).

Let the rotational and the translation motion of the cameras be functions of the variable $t$. Since every continuous function can be approximated using Taylor series, without loss of generality, we assume that these velocities are polynomials. In particular, the rotational velocities are polynomials of order $n_l$ and $n_r$ and the translational velocities are polynomials of order $m_l$ and $m_r$ for the left and right cameras respectively. Under these motion models, using equation (12) one can show that the TFM is a matrix function which is a polynomial of order:

$$\deg \tilde{\mathcal{F}}_{i,j}(t) = \max(n_l, n_r, m_l, m_r) + 1. \qquad (13)$$

Next, we will discuss the action matching criteria which based on the dynamic scene geometry sketched above.

### 3.2. On the Similarity of Two Actions

The videos of actions performed by two actors may not appear similar if they are captured by moving cameras at different times in different settings. For instance, a landmark point which is stationary during the execution of an action, may appear moving due to the camera motion. This is illustrated in Fig. 5a and b where trajectories of an actor performing the picking up action captured from the same viewpoint but with different camera motions are shown. As seen from the figure the trajectories of stationary points, e.g. feet, appear different.

It is known that for the "uncalibrated stationary cameras", there exists a unique fundamental matrix which can be determined from a set of corresponding points (landmark points on the actors) in two views [9, Thr. 9.1]. However, this theorem is not valid for the moving cameras. For the action matching problem in the moving cameras, we can extend the above theorem by using the dynamic epipolar geometry proposed in section 3.1.2. Thus, we have the following:

**Proposition 1** *Given video sequences captured by two uncalibrated moving cameras, there exists a unique temporal fundamental matrix which can be computed using a set of corresponding landmark points on the actors.*

Let an action be represented by the action matrix discussed in Section 2. In particular, the action matrix $\mathbf{U}$ can be organized into

$$\mathbf{U} = (\mathbf{U}_1|\mathbf{U}_2|\dots|\mathbf{U}_n)^\top, \qquad (14)$$

where $\mathbf{U}_t$ is a $3\times 13$ matrix which contains all the landmark points at time $t$. For two actions represented by $\mathbf{U}^{one}$ and $\mathbf{U}^{two}$, based on Proposition 1, and using the polynomial camera motion models, the following holds:

$$\mathbf{U}_t^{one\top}\left(\sum_{i=0}^{k}\mathcal{F}_i t^k\right)\mathbf{U}_t^{two} = 0 \qquad (15)$$

where $\mathcal{F}_i$ is the $3\times 3$ coefficient matrix of the $k^{th}$ order temporal fundamental matrix[7]. Note that, for each time instant only $t$ changes and $\mathcal{F}_i$ remains constant.

So far, we have discussed the existence of a relation (equation (15)) between video sequences captured by two uncalibrated moving cameras. Given the action matrices of two actions, we need to define a similarity metric to quantitatively evaluate how similar the actions are. The only unknowns of equation (15) are the elements in the temporal fundamental matrix, $\mathcal{F}_i$. For saving space, without loss of

[6] In this paper, the intrinsic camera parameters are assumed to be constant, practically during the action focal length does not change.

[7] Order $k$ is given in equation (13) and discussed therein.

generality let us assume the temporal fundamental matrix is a polynomial matrix function of order two, such that we have twenty seven unknowns. Writing equation (15) as a system of linear equations, we have,

$$\mathcal{M}\mathbf{f} = \begin{pmatrix}\mathcal{M}_1^\top & \mathcal{M}_2^\top & \dots & \mathcal{M}_n^\top\end{pmatrix}^\top \mathbf{f} = 0, \qquad (16)$$

where,

$$\mathcal{M}_t = \begin{pmatrix} x_1x_1' & x_1y_1' & x_1 & y_1x_1' & y_1y_1' & y_1 \\ x_1'y_1' & 1 & x_1x_1't & x_1y_1't & x_1t & y_1x_1't \\ y_1y_1't & y_1t & x_1't & y_1't & t & x_1x_1't^2 \\ x_1y_1't^2 & x_1t^2 & y_1x_1't^2 & y_1y_1't^2 & y_1t^2 & x_1't^2 \\ y_1't^2 & t^2 & & & & \end{pmatrix}$$

and $\mathbf{f} = (\mathcal{F}_{1,1}|\ \mathcal{F}_{1,1}|\ \mathcal{F}_{1,1}|\ \mathcal{F}_{2,1}|\ \mathcal{F}_{2,2}|\ \mathcal{F}_{2,3}|\ \mathcal{F}_{3,1}|\ \mathcal{F}_{3,2}|\ \mathcal{F}_{3,3})^\top$, where $\mathcal{F}_{i,j}$ denotes $i^{th}$ coefficient matrix and $j^{th}$ row. Matrix $\mathcal{M}$ is a $13n \times 27$, and assuming the existence of a non-zero solution, $\mathcal{M}$ must be rank deficient, i.e. for $n \geq 27$ rank of $\mathcal{M}$ is at most 26. The solution of $\mathbf{f}$ is given by the unit eigenvector of the covariance matrix $\mathcal{M}^\top\mathcal{M}$ corresponding to the smallest eigenvalue. Once $\mathbf{f}$ is estimated, for a given time instant $t$, we can compute the temporal fundamental matrix by imposing the rank two constraint using SVD [8].

The similarity between two actions can be computed from the quality of the recovered geometry. We use two different criteria to evaluate the quality of the recovered geometry. The first criterion measures how well-conditioned the homogenous system given in equation (16) is. This is done by computing the condition number[8] $\mathcal{C}$ of $\mathcal{M}^\top\mathcal{M}$. For a well-conditioned equation system, the condition number is at infinity. However, in our case, due to noise in the observations, it will not be at infinity. Computing the condition number alone does not guarantee correct estimation of the multi-view geometry [18]. This is mainly due to the existence of multiple solutions of $\mathbf{f}$ in equation (16). However, in the multi-view geometry, there exists a unique solution. Following this observation, we use a second criterion which evaluates the quality of the recovered geometry by computing the symmetric epipolar distance. Symmetric epipolar distance is the average distance of each point in the left (right) camera view to the epipolar line generated from the corresponding point in the right (left) camera view using the estimated temporal fundamental matrix:

$$\mathcal{G} = \sqrt{\left(\frac{x^\top(t)u_l(t)}{|u_l(t)|}\right)^2 + \left(\frac{x'^\top(t)u_r(t)}{|u_r(t)|}\right)^2}, \qquad (17)$$

where $|.|$ denotes norm 2, and $u_l(t) = \tilde{\mathcal{F}}^\top(t)x'(t)$ and $u_r(t) = \tilde{\mathcal{F}}(t)x(t)$ are the epipolar lines at time $t$ corresponding to the point $x(t)$ in the left view and the point $x'(t)$ in the right view. Given these two quantitative measures on the matching of actions, we have the following three cases:

1. The condition number of the covariance matrix is very low, which means that the equation system given in (16) is ill-conditioned. In this case, two actions are declared as different.
2. The condition number is high, however the recovered per frame geometry is ambiguous[9], such that symmetric epipolar distances are high. Similar to the previous case, the actions are declared as different.
3. The condition number is high and the symmetric error is low, this case indicates that the two actions match.

These two quantities can be unified into a single similarity metric, $\mathcal{S} = \big(1 - \exp\big(-\frac{\mathcal{C}^2}{\sigma_c^2}\big)\big)\big(\exp\big(-\frac{\mathcal{G}^2}{\sigma_g^2}\big)\big)$, where $\sigma_c$ and $\sigma_g$ controls the allowed range of changes in $\mathcal{C}$ and $\mathcal{G}$ respectively. During our experiments, we fixed the values of both $\sigma_c$ and $\sigma_g$. Empirical justification of the similarity metric will be given in the next section.

## 4. Experiments

To validate the proposed matching approach, we performed a set of experiments on two different applications. The first set of experiments involved recognition of an action from a database of known actions. Since there is no standard database of action videos captured using moving cameras, we generated our own database of eighteen different actions. The actions are performed by different actors in different environments. In Fig. 6, we show the complete set of actions in our action database. Since the actors appear quite small in the video, the labeled landmark points are quite noisy. Thus in addition to the camera motion, the noisy landmarks makes the recognition task harder. The second set of experiments involved the retrieval of an exemplar action from a long video.

### 4.1. Action Recognition

Using the complete set of actions in Fig. 6, we compute the matching score using similarity metric between each action with every other action. The results are demonstrated by a confusion matrix given in Fig. 7a. In the figure black illustrates similar actions and white illustrates dissimilar actions. For all the actions, the action categories are correctly clustered. An unimportant clustering is observed for two sitting down actions, which are confused with one of the walking actions. The main reason of the confusion is that the particular actor performing the walking action is quite

[8]Condition number is the ratio between the maximum and the minimum singular value and measures the worst-case loss of precision of a linear system.

[9]Note that, after computing the temporal fundamental matrix unknowns, we can estimate per frame fundamental matrices by substituting the time $t$.

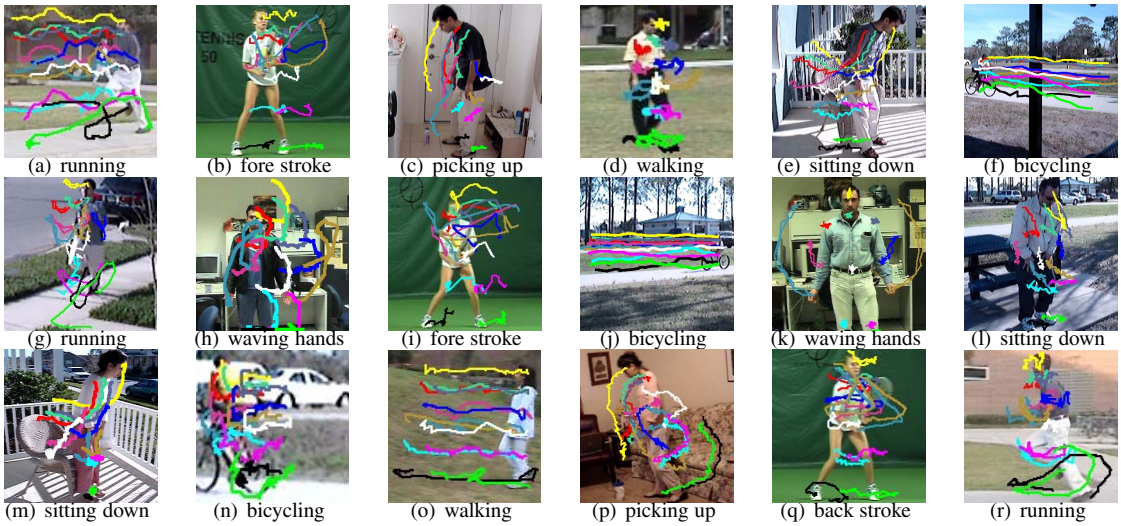
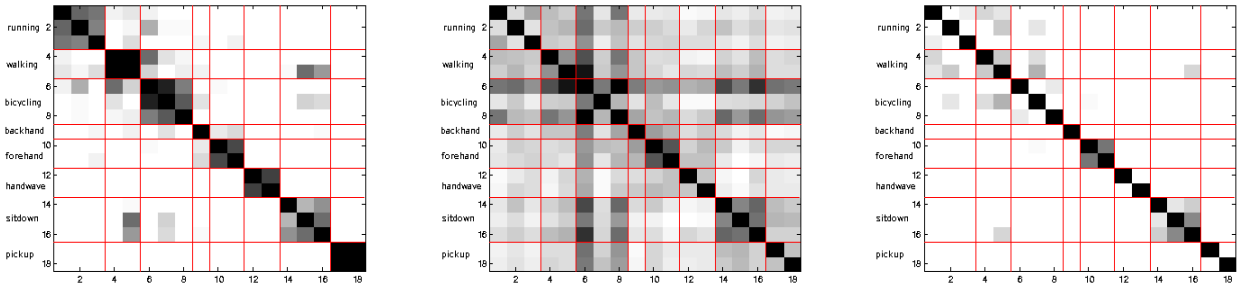



**Figure 6. Set of actions used to test the proposed method. For each action, we display the first image of the sequence along with the trajectories of the landmark points superimposed.**




(a) (b) (c)

**Figure 7. Confusion matrices computed for three different methods. The clusters boundaries are indicated by red lines. The dark color in small squares indicates high action similarity and the light color indicates low similarity. (a) Proposed recognition method using dynamic epipolar geometry. Note that, except for a few outliers, clusters are correctly found. (b) Static epipolar geometry, where the similarity between the actions is computed using the condition number alone. (c) Static epipolar geometry, where both the condition number and the symmetric epipolar distance are used.**

small in size therefore even one pixel error in the location of landmark points is intolerable.

In Fig. 7, we compare the performance of the proposed action recognition approach (part (a))with the static epipolar geometry based approaches (parts (b) and (c)). The images show the confusion matrices for all three methods. Two variants of static epipolar geometry based approaches is shown. In part (b), we directly used the condition number which was shown to be successful for the recognition task in the previous approaches; in part (c), we show the results of static epipolar geometry and employing the proposed similarity metric. It is qualitatively evident from the figures that the proposed approach correctly finds the action clusters.

### 4.2. Action Retrieval

The task in this experiment is to retrieve the occurrences of a particular action in a long video. We used a long tennis sequence, in which a tennis player is performing various actions, such as forehand stroke, walking, etc. In particular, given exemplars of tennis stroke and walking actions, we attempted to retrieve the occurrences of these exemplars throughout the tennis video. In Fig. 8, we demonstrate the performance both qualitatively and quantitatively. On the

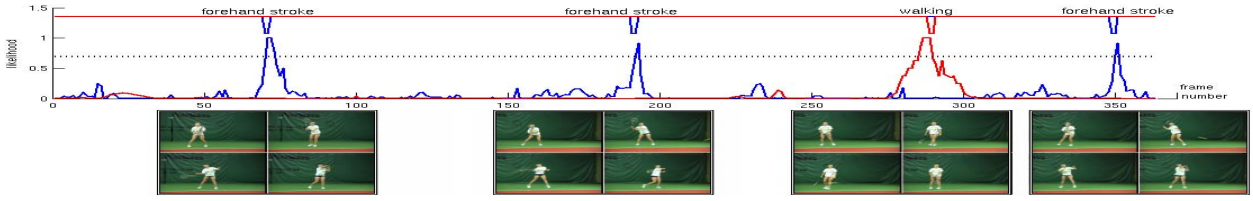



**Figure 8. Quantitative and qualitative evaluation of the action retrieval application in a long tennis sequence. The plot on the top shows the similarity metric computed for the walking action (red plot) and the forehand stroke (blue plot) along with the ground truth data shown on the top of the plot. On the bottom, we show the set of corresponding images of the retrieved actions.**

top, we show the similarity metric along with the ground truth marked as delta on the top of the plot. The blue plot indicates the similarity of tennis stroke action and the red plot indicates the similarity of walking action. On the bottom of the figure, we show a set of corresponding images of the retrieved actions. As it is clear from the figure there is very obvious peak corresponding to the correct match.

## 5. Conclusion

We proposed a novel approach for recognition of human actions in videos captured by moving cameras. Proposed approach uses the geometry of dynamic scenes, which is another contribution of this paper. The proposed approach is demonstrated to perform robust recognition and retrieval of actions in a number of challenging sequences, which contain different views (moving camera) of the same action performed by different actors in different environments.

**Acknowledgments:** This material is based upon work funded in part by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

## References

[1] S. Avidan and A. Shashua. Threading fundamental matrices. *PAMI*, 23(1), 2001.

[2] M. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *IJCV*, 25(1):23–48, 1997.

[3] A. Bobick and J. Davis. The representation and recognition of action using temporal templates. *PAMI*, 23(3):257–267, 2001.

[4] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003.

[5] O. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig. In *ECCV*, pages 563–578, 1992.

[6] A. Fitzgibbon and A. Zisserman. Multibody structure and motion: 3-d reconstruction of independently moving objects. In *ECCV*, 2000.

[7] A. Gritai, Y. Sheikh, and M. Shah. On the invariant analysis of human actions. In *ICPR*, 2004.

[8] R. Hartley. In defence of the 8-point algorithm. In *CVPR*, page 1064 1070, 1995.

[9] R. Hartley and A. Zisserman. *Multiple View Geometry in computer Vision-second edition*. Cambridge Un. Press, 2004.

[10] H. Ishikawa and D. Geiger. Occlusions, discontinuities, and epipolar lines in stereo. In *ECCV*, 1998.

[11] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 73(2):201–211, 1973.

[12] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[13] V. Parameswaran and R. Chellappa. Quasi-invariants for human action representation and recognition. In *ICPR*, volume 1, pages 307–310, 2002.

[14] R. Polana and R. Nelson. Recognizing activities. In *CVPR*, 1994.

[15] C. Rao, A. Gritai, M. Shah, and T. Syeda-Mahmood. View-invariant alignment and matching of video sequences. In *ICCV*, 2003.

[16] C. Rao, A. Yilmaz, and M. Shah. View invariant representation and recognition of actions. *IJCV*, 50(2):203–226, 2002.

[17] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *Motion- Based Recognition*. Kluwer, 1996.

[18] P. Torr and A. Zisserman. Performance characterisation of fundamental matrix estimation under image degradation. *MVA Jrn.*, 9:321–333, 1997.

[19] E. Trucco and A. Verri. *Introductory techniques for 3d computer vision*. Prentice Hall, 1998.

[20] L. Wolf and A. Shashua. On projection matrices pk–p2, and their applications in computer vision. *IJCV*, 48(1):53–67, 2002.

[21] M. Yang, N. Ahuja, and M. Tabb. Extraction of 2d motion trajectories and its application to hand gesture recognition. *PAMI*, 24(8):1061–1074, 2002.

[22] A. Yilmaz and M. Shah. Action sketch: A novel action representation. In *CVPR*, 2005.

[23] Z. Zhang. An automatic and robust algorithm for determining motion and structure from two perspective images. In *Int. Conf. on Computer Analysis of Images and Patterns*, pages 174–181, 1995.