Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 104 (2006) 221–231
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Matching actions in presence of camera motion

Alper Yilmaz [a,*], Mubarak Shah [b]

[a] Department of CEEGS, Ohio State University, Columbus OH-43035, USA
[b] School of EECS, University of Central Florida, Orlando FL-32828, USA

Received 15 February 2006; accepted 25 July 2006
Available online 27 September 2006

Abstract

When the camera viewing an action is moving, the motion observed in the video not only contains the motion of the actor but also the motion of the camera. At each time instant, in addition to the camera motion, a different view of the action is observed. In this paper, we propose a novel method to perform action recognition in presence of camera motion. Proposed method is based on the epipolar geometry between any two views. However, instead of relating two static views using the standard fundamental matrix, we model the motions of independently moving cameras in the equations governing the epipolar geometry and derive a new relation which is referred to as the "temporal fundamental matrix." Using the temporal fundamental matrix, a matching score between two actions is computed by evaluating the quality of the recovered geometry. We demonstrate the versatility of the proposed approach for action recognition in a number of challenging sequences.

Keywords: Action recognition; Epipolar geometry; Video; Action retrieval; Moving camera

1. Introduction

Over the last decade, action recognition has received significant importance due its applications in human–computer interaction, event based video retrieval, and automated surveillance. In particular, action recognition is complex when two instances of the same action are performed by two different actors. The complexity increases in manifolds if these actions are viewed from different camera viewpoints. In order to simplify the action recognition problem, a common strategy adopted by researchers is to assume the unknown action and its corresponding action in the database share the same camera view. For instance, if the exemplar walking action is fronto parallel, then the input action to be matched has to be fronto parallel. Using this formalism, Efros et al. [1] match two actions by computing the normalized correlation between a set of features extracted from the optical flow. Similarly, Polana and Nelson [2] use template matching in the spatio–temporal space, where the templates are generated from the statistics of normal flow. An alternative to the optical flow is to use the spatial and temporal image derivatives. Manor and Irani [3] compare empirical distributions of two actions which are generated from the temporal image gradient at various temporal scales. Similarly, Laptev and Lindeberg [4] match representative points extracted from spatial and temporal image gradients. This is essentially an extension of the Harris corner detector [5] to the spatio–temporal space. Action recognition can also be performed by computing the affine motion between the corresponding segments in two consecutive frames. Following this strategy, Yang et al. [6] generate an affine trajectory of the hand and head regions for sign language recognition. Matching score between two actions is computed using a time delay neural network. Black and Yacoob [7] compare the rate of change of the affine parameters that are computed from the bounding boxes around the eyes, eyebrows and the mouth. The approach proposed by Bobick and Davis [8] evaluates the Mahalanobis distance between the Hu moments computed from the motion history templates extracted from a stack

* Corresponding author.
E-mail addresses: yilmaz.15@osu.edu (A. Yilmaz), shah@cs.ucf.edu (M. Shah).

1077-3142/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2006.07.012

of silhouettes. Blank et al. [9] also utilize a stack of silhouettes to compute spatio–temporal properties, which are extracted by evaluating a Poisson equation for each point in the silhouette stack. Given these features, a matching score between two actions is computed by evaluating the Euclidean distance between the features.

However, fixing the camera viewpoint and considering only one view of the action is not suitable for solving practical problems that may arise during a surveillance scenario where an action can be observed from different viewing angles. Therefore, "view invariance" has become very important in computer vision. An intuitive solution to deal with the viewpoint changes is to set up an environment with multiple cameras. Multiple camera have also been successfully applied in the motion capture work [10,11]. For action recognition using multiple cameras, Weinland et al. [12] computed the Fourier descriptors from the motion history volumes generated from a stack of silhouettes for each view. Despite its simplicity, use of multiple cameras is generally impractical and expensive. A common approach to achieve view invariance is to employ the fundamental matrix between two different views of an action. In order to achieve this, trajectories of the body parts [13,14] and the actor silhouettes [15,16] have been used. Rao et al. [17] compute a matching score between two actions by evaluating the fundamental matrix using the selected points (points corresponding to maxima in spatio–temporal curvature) on two trajectories. Syeda-Mahmood et al. [15] extracted eight matching features from the silhouettes of two input actions to compute a matching score using the fundamental matrix. An alternative approach to achieve view invariance is to extract view invariant features from the video. In [18], Parameswaran et al. conjectured that during the course of an action, when five landmark points on the human body form a plane in 3D, a set of projective invariants can be extracted in the image domain and used for matching actions.

Although above methods can match two different views of an action, they rely on stationary cameras. Thus, use of *stationary cameras* has become a defacto assumption in the action recognition literature. In many videos such as broadcast news, movies, UAV (Unmanned Air Vehicles) videos, and home videos the stationary camera assumption is violated. Consider videos clips about the same event broadcasted by different news networks. The videos may look very different from each other due to the different camera viewpoints and motions. Similarly, a home video of the same event captured by a stationary person cannot be matched with a video captured by a person who is moving around as he captures the video. Also, it is difficulty to match videos captured by two UAVs from different viewpoints and with different motions. This complexity is mainly due to the camera motion which induces a global motion in the video in addition to the actor's motion during performance of an action. Additionally, due to the camera motion different views of the actor may be observed at each time instant.

A possible approach to matching actions in presence of camera motion is to use the standard fundamental matrix. However, standard fundamental matrix is valid for relatively stationary cameras[1] [19], hence, we have to treat each frame independently and compute instantaneous matching scores using each fundamental matrix. This approach, however, is not attractive for a number of reasons. First and foremost, the computation of the fundamental matrix (eight point algorithm) requires at least eight visible point correspondences at each time instant. Due to self occlusion of the body parts this constraint may not be met throughout the course of an action. In addition, since every frame has an independent set of fundamental matrix parameters, number of required correspondences linearly increase with the number of frames. For instance for an action of 30 frames, $8 \times 30 = 240$ point correspondences are required. Third, estimating independent fundamental matrices from a sequence, in which consecutive frames are related to each other by the camera motion, requires additional constraints, such as trifocal tensor, to enforce temporal dependency [20]. Last but not least, instantaneous constraints (use of observations on a frame basis) to match actions are weaker constraints than using the observations all together to find a global solution.

Another possible approach to perform matching in moving camera setting is to compensate the camera motion using planar homography. However, this is only possible for distant views and planar scenes, which is too restrictive in the context of actions. Recently, there has been a body of research dealing with moving cameras for the structure from motion problem [21,19]. However, these methods require rigid objects, hence, they are not applicable to the action recognition.

This paper proposes a novel approach to perform action recognition in presence of camera motion. Proposed method makes use of the epipolar geometry between two cameras. In contrast to the standard fundamental matrix derived from the epipolar geometry, we model the rotational and translational camera motions and derive a new relation between two camera views. This new relation, which is referred to as the "temporal fundamental matrix" (TFM), is in the form of a matrix function. To compute a matching score between two actions, we first hypothesize that the actions satisfy TFM and test the hypothesis by evaluating the recovered geometry. The experiments performed for three different applications show the robustness of TFM in presence of camera motion. The applications include: action recognition, action retrieval, and object association across cameras.

The paper is organized as follows. In Section 2, we discuss the action representation used in our appaorch. Section 3 describes the proposed approach to relate two actions viewed from moving cameras. The experiments

[1] Relatively stationary cameras are defined as two cameras with zero motion or two cameras mounted on a stereo rig and moving together.

demonstrating the performance of the proposed approach is given in Section 4. Subsequently we present our conclusion in Section 5.

2. Representing human actions

Actions are intensional space–time events performed by moving the body parts in a coordinated fashion [23]. An ideal representation of this space–time event would be in a 4D space: (X, Y, Z, t) in terms of the trajectories Γ of all the points on the actor. Although, it is desirable to work in the 4D space, video contains only the spatio–temporal projection (x, y, t) of this space–time event. In his seminal work, Johansson [22] shows that only a set of few 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. In this paper, we follow this line of work and represent an action as a set of trajectories obtained from tracking the landmark points on the human body (see Fig. 1). This type of action representation is also supported by a large body of research work on articulated object tracking such as [24,25], which has been discussed and categorized in the surveys by Aggarwal and Cai [26], Gavrilla [27], and Moeslund and Granum [28]. Tracking of landmark points results in a set of spatio–temporal trajectories $\Gamma_i = (x_1, x_2, \ldots, x_n)^\top$, where i is the label of the landmark points, n is the duration of the action and $x = (xy1)^\top$. Given these trajectories, we represent an action by a collection of thirteen trajectories:

$$U = (\Gamma_1^\top, \Gamma_2^\top \ldots \Gamma_{13}^\top), \tag{1}$$

that we call the "action matrix", which is a $3 \times 13n$ matrix. In Fig. 2a, we show several frames from a walking sequence. Fig. 2b shows the action trajectories in the spatio–temporal space.

In the rest of the paper, we assume tracking has already been performed and we are given a set of joint trajectories. However, human body joint tracking in an unconstrained environment is quiet complex and is an active topic of current research in computer vision. There are two important issues related to joint tracking: occlusion and accuracy. Some joints may be occluded during the motion of an actor or camera from a particular viewpoint. Since we are using a redundant set of joints (thirteen or less joints) in several frames, it is not a problem if a few joints are occluded during some number of frames. Since we use several frames and employ temporal information to distinguish between different actions, exact locations of the joints in a particular frame are not that crucial. Our experiments show that we are able to easily distinguish similar actions, e.g., walking and running based on very noisy tracks.

3. Theory and analysis of matching actions

When the camera viewing an action is moving, the motion observed in the video not only contains the local actor motion but also the camera motion. Due to this, motion trajectories [29], silhouettes [8], optical flow vectors [1] or image derivatives [3] extracted from the video without global motion compensation cannot uniquely characterize an action. In this section, we will present our approach for matching actions in presence of camera motion without motion compensation.

Before discussing matching of actions in the moving camera setting, we first discuss stationary camera setting. In this setting, a matching score between the two actions can be computed based on epipolar geometry [13–16]. Epipolar geometry is defined by projecting a world point $\mathbf{P} = (X, Y, Z)$ to the left camera reference frame, $\mathbf{P}_l = \mathbf{R}_l\mathbf{P} + \mathbf{T}_l$, and the right camera reference frame, $\mathbf{P}_r = \mathbf{R}_r\mathbf{P} + \mathbf{T}_r$ (see Fig. 3), which are related by :

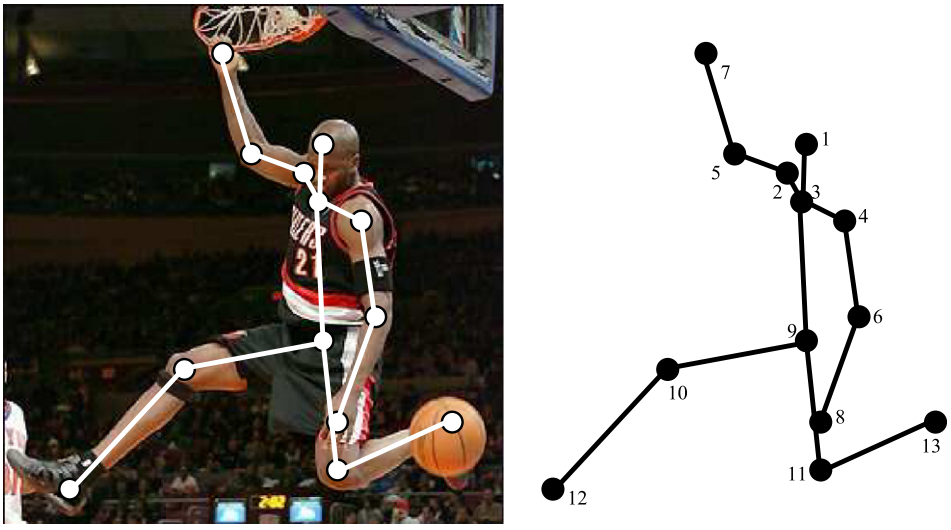


Fig. 1. Point-based representation of an actor. Johansson [22] has shown that this representation provides sufficient information to perceive an action performed by an actor.

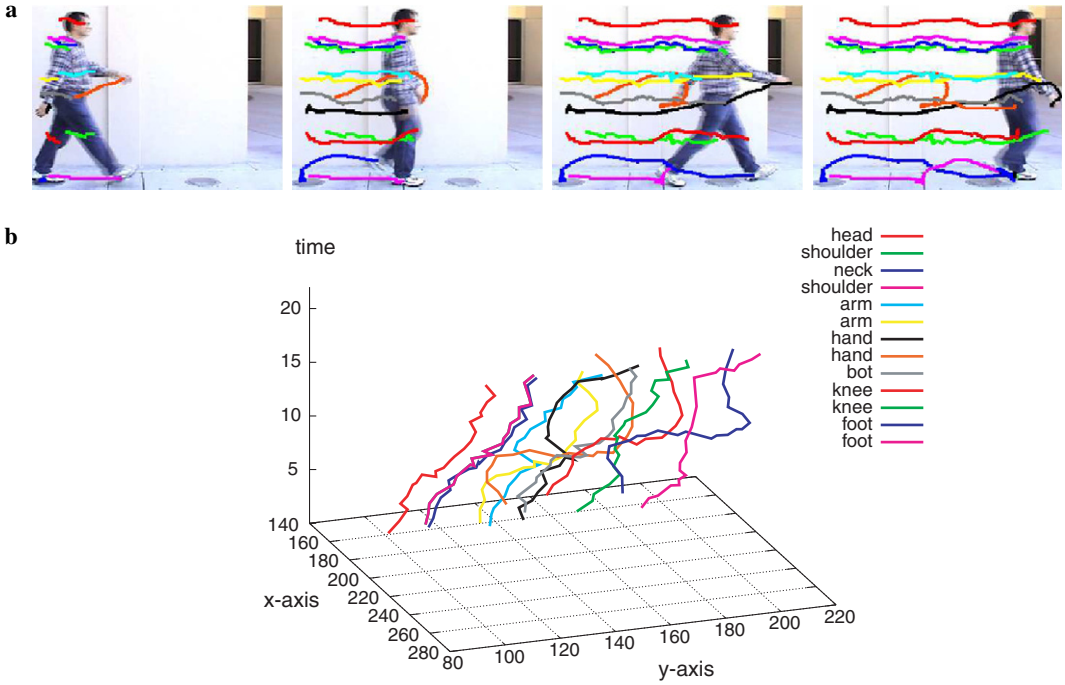


Fig. 2. Trajectories of 13 landmark points of a walking person (a) superimposed on the images, and (b) shown in the spatio–temporal space. The sequence is taken from a stationary camera.

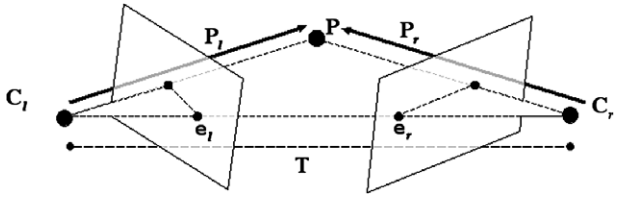


Fig. 3. Epipolar geometry defined for two cameras observing the same scene. **P** denotes a 3D point (a landmark point on the actor performing the action), $\mathbf{C}_l$ and $\mathbf{C}_r$ are left and right camera centers, $\mathbf{e}_l$ and $\mathbf{e}_r$ are the epipoles of the left and right image planes.

$$\mathbf{P}_r = \underbrace{\left(\mathbf{R}_r\mathbf{R}_l^\top\right)}_{\mathbf{R}:\text{relativerotation}} \mathbf{P}_l - \underbrace{\left(\mathbf{R}_r\mathbf{R}_l^\top\mathbf{T}_l - \mathbf{T}_r\right)}_{\mathbf{T}:\text{relativetranslation}}. \tag{2}$$

This relation along with coplanarity condition defined for the epipolar plane result in the well-known formula:

$$\mathbf{P}_r^\top(\mathbf{RS})\mathbf{P}_l = \mathbf{P}_r^\top\mathscr{E}\mathbf{P}_l = 0, \tag{3}$$

where $\mathbf{S}$ is a rank deficient matrix obtained from relative translation $\mathbf{T}$, and $\mathscr{E}$ is the essential matrix [30]. The essential matrix $\mathscr{E}$ 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:

$$\mathbf{x}_r^\top(M_r^{-\top}\mathscr{E}M_l^{-1})\mathbf{x}_l = \mathbf{x}_r^\top\mathscr{F}\mathbf{x}_l = 0, \tag{4}$$

where $\mathscr{F}$ is a 3×3 matrix referred to as the fundamental matrix [31].

A common approach to compute a matching score between two actions using the fundamental matrix is based on some measure defined from the observation matrix $\mathbf{O}$ given by:

$$\mathbf{Of} = \begin{pmatrix} x_{r,1}x_{l,1} & x_{r,1}y_{l,1} & x_{r,1} & y_{r,1}x_{l,1} & y_{r,1}y_{l,1} & y_{r,1} & x_{l,1} & y_{l,1} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{r,m}x_{l,m} & x_{r,m}y_{l,m} & x_{r,m} & y_{r,m}x_{l,m} & y_{r,m}y_{l,m} & y_{r,m} & x_{l,m} & y_{l,m} & 1 \end{pmatrix}\mathbf{f} = 0, \tag{5}$$

where $\mathbf{f} = (\mathscr{F}_{11}, \mathscr{F}_{12}, \mathscr{F}_{13}, \mathscr{F}_{21}, \mathscr{F}_{22}, \mathscr{F}_{13}, \mathscr{F}_{31}, \mathscr{F}_{32}, \mathscr{F}_{33})^\top$, and m is the number of corresponding points across views [32]. The system of equations given in Eq. (5) is homogeneous and has a rank of 8. In an ideal case, where there are no observation errors and two actions match, the ninth singular value equals to zero, such that the ninth singular value can be used to compute matching score. Another possibility is to consider the condition number of $\mathbf{O}$, which is given by the ratio between the first and the ninth singular value. In case of matching actions based on the condition number, minimum score implies dissimilar actions.

Consider Fig. 4 where trajectories of the landmark points appear very different for two actors performing the same picking up action. Therefore, in the case of moving cameras, the matching scores described using the standard fundamental matrix cannot be used for action recognition. This is due to the changing epipole positions at each time instant which results in a new geometric relation (see Fig. 5).

In order to analyze the effect of camera motion on the geometric relation defined for two views of an action, we introduce the rotational and translational camera motions in the derivation of the essential and fundamental matrices. Particularly, Eq. (2) becomes $\mathbf{P}_r(t) = \mathbf{R}(t)\ \mathbf{P}_l(t) - \mathbf{T}(t)$, where every component of the matrix is a function of time t, hence, we have the following observation:

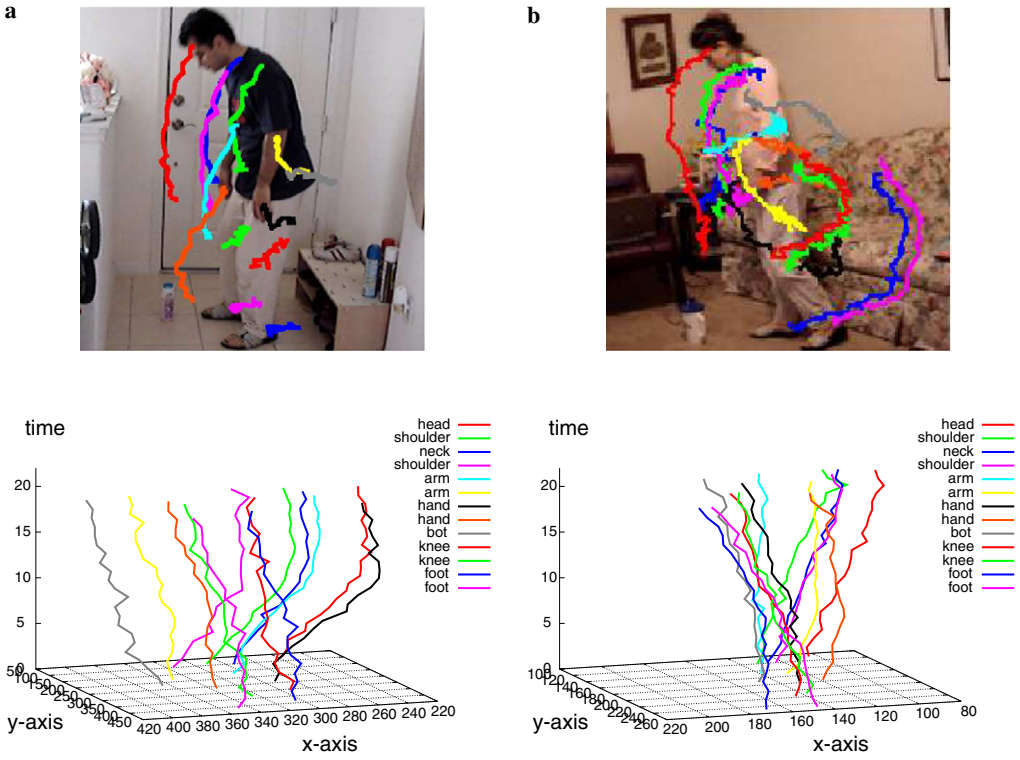


Fig. 4. Trajectories of the landmark points of two actors performing the picking up action which is captured from different viewpoints. In both sequences, cameras undergo different motions. As observed, the trajectories are not similar. (a) Camera is translating in the x axis, (b) camera is rotating and translating around the z axis.

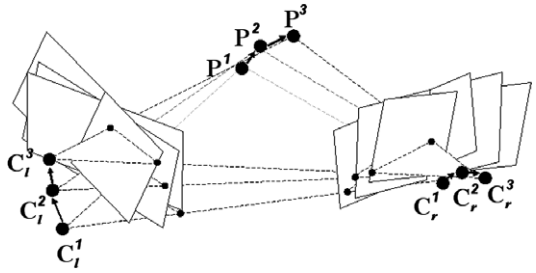


Fig. 5. Epipolar geometry between two views for moving cameras. At each time instant, the locations of the epipoles change which results in a new fundamental matrix between the views.

Observation 1 (Temporal fundamental matrix)**.** *For two independently moving cameras, the pixels locations at time t in the left and right camera image planes are related to each other through a time varying matrix function of the form*:

$$\mathbf{x}_r^\top(t)\underbrace{M_r^{-\top}\overbrace{\mathbf{R}(t)\mathbf{S}(t)}^{\mathcal{E}(t)}M_l^{-1}}_{\mathcal{F}(t)}\mathbf{x}_l(t)=0, \tag{6}$$

where $\tilde{\mathcal{F}}(t)$ is referred to as the temporal fundamental matrix (TFM).

Under this observation, each component of the TFM is a function of time, such that we have $f_{11}(t)$, $f_{12}(t)$, $f_{13}(t)$, $f_{21}(t),\ldots,f_{33}(t)$. Assuming that the internal camera parameters do not change, each component $f_{ij}(t)$ relies only on translational and rotational camera motions.

We use video clips of temporal length of two to three seconds for action matching. In this time period, motion of the cameras can be approximated by polynomial functions. In particular, let the rotational motion be polynomials of order n_l and n_r and the translational motion be polynomials of order m_l and m_r for the left and right cameras respectively. Under these definitions, TFM has the following form:

Theorem 1 (Small motion constraint)**.** *For relatively moving cameras, the temporal fundamental matrix $\tilde{\mathcal{F}}(t)$ is a 3×3 matrix whose components are polynomials of order*:

$$\deg \tilde{\mathcal{F}}_{ij}(t)=\max\left((n_{lr}+m_l),(n_{lr}+m_r)\right), \tag{7}$$

where $n_{lr}=\max(n_l, n_r)$.

In order to prove the lower bound of this theorem, we first discuss the degree of polynomials (DOP) in the camera projection matrices, then we will provide DOP for relative rotation and relative translation and finally, the degree of polynomials in TFM. Under the small motion constraint, we first define rotational and translational motion of the left camera (the following discussion applies to the right camera as well)[2].

[2] Due to small motion constraint, we can assume that the first order Taylor series expansion of the rotational velocities is adequate such that $\cos\alpha=1$ and $\sin\alpha=\alpha$.

$$\Delta\Omega_l(t) = \begin{bmatrix} 1 & -\omega_z(t) & \omega_y(t) \\ \omega_z(t) & 1 & -\omega_x(t) \\ -\omega_y(t) & \omega_x(t) & 1 \end{bmatrix}, \quad \Delta\Theta_l(t) = \begin{bmatrix} \theta_x(t) \\ \theta_y(t) \\ \theta_z(t) \end{bmatrix}. \tag{8}$$

Since the camera pose changes at each time instant, camera projection matrix becomes:

$$\mathbf{P}_l(t) = \underbrace{\prod_{i=0}^{t} \Delta\Omega_l(i)}_{\Omega_l(t)} \mathbf{P}_{\text{world}} + \underbrace{\sum_{i=0}^{t} \Delta\Theta_l(i)}_{\Theta_l(t)}, \tag{9}$$

where $(\Omega_l(0), \Theta_l(0))$ denote the initial camera pose. The rotational component of the projection matrix, after the cascade of the rotational motions, becomes:

$$\Omega_l(t) = \begin{bmatrix} 1 & -\sum_i \omega_z(i) + \epsilon & \sum_i \omega_y(i) + \epsilon \\ \sum_i \omega_z(i) + \epsilon & 1 & -\sum_i \omega_x(i) + \epsilon \\ -\sum_i \omega_y(i) + \epsilon & \sum_i \omega_x(i) + \epsilon & 1 \end{bmatrix} \Omega_l(0), \tag{10}$$

where ϵ includes terms such as $\sum \omega_i\omega_j$, $\sum \omega_i\omega_j\omega_k$, etc. Due to small angle approximation, ϵ is approximately 0 and can be dropped. By definition, $\omega_x(t)$, $\omega_y(t)$, and $\omega_z(t)$ are polynomials of order n_l. Hence, the DOP in $\Omega_l(t)$ becomes n_l. Similarly, DOP in $\Theta_l(t)$ given in Eq. (9) is m_l. For the moving camera system, relative rotation between the cameras is given by $\mathbf{R}(t) = \Omega_r(t)\Omega_l^{\top}(t)$. Based on Euler angles the DOP in $\mathbf{R}(t)$ can be computed as $\max(n_l, n_r)$. The relative translation between the moving cameras is given by $\mathbf{T}(t) = \mathbf{R}(t)\Theta_r(t) - \Theta_l(t)$. In order to find DOP in $\mathbf{T}(t)$, we first need to consider $\mathbf{R}(t)\Theta_r(t)$. Multiplication of these terms results in summation of the $\deg\mathbf{R}(t)$ and $\deg\Theta_r(t)$. Combining this with the second term in $\mathbf{T}(t)$, DOP in $\mathbf{T}(t)$ becomes $\max(\max(n_l, n_r) + m_l, m_r)$. Due to the multiplication of $\mathbf{R}(t)$ with the rank deficient matrix version of $\mathbf{T}(t)$ to produce $\mathscr{E}(t)$, the DOP of $\mathscr{E}(t)$ and $\mathscr{F}(t)$ is:

$$\deg\mathscr{F}(t) = \deg\mathscr{E}(t) = \max((n_{lr} + m_l), (n_{lr} + m_r)), \tag{11}$$

where $n_{lr} = \max(n_l, n_r)$. Based on this result:

$$\mathscr{F}(t) = \sum_{i=0}^{k} \mathscr{F}_i t^k,$$

where $\mathscr{F}_i$ is the 3×3 coefficient matrix of the kth order TFM stisfying Eq. (6). Note that for stationary cameras, where the polynomial orders are $k = 0$, TFM reduces to the standard fundamental matrix.

Given n corresponding trajectories in two sequences, the coefficients in the polynomials of the TFM can be estimated by rearranging Eq. (6) as a linear system of n equations. Let us assume $\deg\mathscr{F}(t) = 2$, such that we have 27 unknowns. In this case, linear system of equations are given by:

$$\mathscr{M}\mathbf{f} = (\mathscr{M}_1^{\top}\mathscr{M}_2^{\top}\ldots\mathscr{M}_n^{\top})^{\top}\mathbf{f} = 0, \tag{12}$$

where

$$\mathscr{M}_t = (x_1x_1' \;\; x_1y_1' \;\; x_1 \;\; y_1x_1' \;\; y_1y_1' \;\; y_1 \;\; x_1'y_1' \;\; 1 \;\; x_1x_1't \;\; x_1y_1't \;\; x_1t \;\; y_1x_1't \;\; y_1y_1't \;\; y_1t \;\; x_1't \;\; y_1't \;\; t \;\; x_1x_1't^2 \;\; x_1y_1't^2 \;\; x_1t^2 \;\; y_1x_1't^2 \;\; y_1y_1't^2 \;\; y_1t^2 \;\; x_1't^2 \;\; y_1't^2 \;\; t^2),$$

and $\mathbf{f} = (\mathscr{F}_{1,1}|\;\mathscr{F}_{1,1}|\;\mathscr{F}_{1,1}|\;\mathscr{F}_{2,1}|\;\mathscr{F}_{2,2}|\;\mathscr{F}_{2,3}|\;\mathscr{F}_{3,1}|\;\mathscr{F}_{3,2}|\;\mathscr{F}_{3,3})^{\top}$, where $\mathscr{F}_{i,j}$ denotes ith coefficient matrix and jth row. Matrix $\mathscr{M}$ is a $13n\times 27$, and assuming the existence of a non-zero solution, $\mathscr{M}$ must be rank deficient, i.e., for $n \geqslant 27$ rank of $\mathscr{M}$ is at most 26. The solution of $\mathbf{f}$ is given by the unit eigenvector of the covariance matrix $\mathscr{M}^{\top}\mathscr{M}$ corresponding to the smallest eigenvalue. Once $\mathbf{f}$ is estimated, an instance of the fundamental matrix at a given time t can be computed by imposing rank two constraint as discussed in [32].

3.1. Computing the matching score

Consider two actions A and B with associated observations in the form of action matrices $\mathbf{U}^A$ and $\mathbf{U}^B$. Our goal is to find how likely A matches B. For this purpose,we introduce an event $\mathscr{L}$ which occurs when both actions are the same. Hence, a matching score between two actions can be defined by the posterior conditional probability $p(\mathscr{L}|\mathbf{U}^A, \mathbf{U}^B)$. Since the observations are independent (clips are taken at different times and the actors are different) and occurrence of any action is equally likely, application of Bayes' rule to this conditional probability results in:

$$p(\mathscr{L}|\mathbf{U}^A, \mathbf{U}^B) = kp(\mathbf{U}^A|\mathscr{L})p(\mathbf{U}^B|\mathscr{L}), \tag{13}$$

where $k = p(\mathscr{L})p(\mathbf{U}^A)^{-1}p(\mathbf{U}^B)^{-1}$. In Eq. (13), the observation $\mathbf{U}^A$ ($\mathbf{U}^B$) has an associated label A (B) such that the condition $\mathscr{L}$ can be replaced by the label of the other action, B (A):

$$p(\mathscr{L}|\mathbf{U}^A, \mathbf{U}^B) = kp(\mathbf{U}^A|B)p(\mathbf{U}^B|A). \tag{14}$$

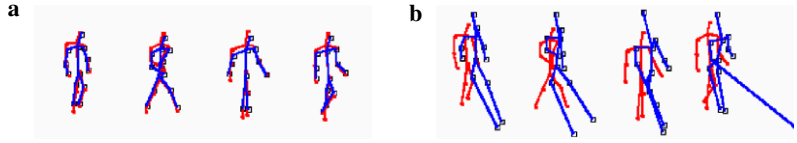
In order to compute these conditional probabilities, we define a relation between $\mathbf{U}^A$ and $\mathbf{U}^B$ based on the aforementioned geometric relation:

Proposition 1. *If two action sequences belong to the same action, trajectories of the corresponding landmark points satisfy a unique polynomial TFM*:

$$\mathbf{U}^{A\top}\mathscr{F}(t)\mathbf{U}^B = 0 \tag{15}$$

where $\mathbf{U}^A$ *and* $\mathbf{U}^B$ *are action matrices* (*Section 2*) *corresponding to each sequence.*

Proposition 1 can be validated by projecting the action viewed in one camera to the other camera. For instance, in the case when both actions are the same, the projected action matrix should be similar to the target action matrix (see Fig. 6). In context of epipolar geometry, projection of a point in one view provides a line in the other view. Given a point in the left image, its corresponding point in the right image can easily be obtained by projecting the location of left point in the right image and finding its closest point on the epipolar line.



Projection of walking to walking Projection of pickup to walking

Fig. 6. (a) Set of representative frames showing the projection of an action (blue) onto the corresponding action from another view (red), (b) same as (a) but this time we project the pickup action to the walking action. As observed, non-matching action does not project as well as the matching action. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

Based on this observation, the similarity between two actions is computed from the quality of the estimated the TFM. Using TFM, for each point x in action matrix $\mathbf{U}^B$, we compute its distance d from the epipolar line u^A generated from the corresponding point x' in action matrix $\mathbf{U}^A$. Summation of all these distances results in the similarity between the actions:

$$\mathrm{d}(\mathbf{U}^A, \mathbf{U}^B) = \sum_{x_j \in \mathbf{U}^B} \frac{x_j^\top(t) u_j^A(t)}{|u_j^A(t)|}, \tag{16}$$

where $|.|$ denotes norm 2, t is frame number and $u_j^A(t) = \mathscr{F}^\top(t) x_j'(t)$, $x_j'(t) \in \mathbf{U}_i^A$. The distance given in Eq. (16) is ideally zero, however, due to noise in the observations it will be greater than zero. Assuming that the observation noise is Gaussianly distributed, the conditional probability $p(\mathbf{U}^A|B)$ becomes:

$$p(\mathbf{U}^A|B) = \frac{1}{2\pi\sqrt{\sigma}} \exp - \frac{\mathrm{d}(\mathbf{U}^A, \mathbf{U}^B)^2}{2\sigma^2}. \tag{17}$$

At this point, we should note that $\mathrm{d}(\mathbf{U}^A, \mathbf{U}^B) \neq \mathrm{d}(\mathbf{U}^B, \mathbf{U}^A)$ such that $p(\mathbf{U}^A|B)$ is different from $p(\mathbf{U}^B|A)$. Substituting these conditional probabilities into Eq. (14), a matching score between two actions can be computed as:

$$p(\mathscr{L}|\mathbf{U}^A, \mathbf{U}^B) = k' \exp - \frac{\mathrm{d}(\mathbf{U}^A, \mathbf{U}^B)^2 + \mathrm{d}(\mathbf{U}^B, \mathbf{U}^A)^2}{2\sigma^2}, \tag{18}$$

where k' is constant. The proposed matching score is related to the symmetric epipolar distance. In particular, taking the negative log likelihood of Eq. (18) results in $\mathrm{d}(\mathbf{U}^A, \mathbf{U}^B)^2 + \mathrm{d}(\mathbf{U}^B, \mathbf{U}^A)^2 + \epsilon$.

4. Experiments

To validate the proposed matching approach, we performed a set of experiments for three 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. 7, we show the complete set of actions in our action database. For most of the actions, the landmark trajectories are noisy. In addition to the noisy trajectories, the motion of the camera makes the recognition task harder. The second set of experiments involved the retrieval of two exemplar actions from a long video where the exemplar actions are performed by actors different from the actor in the long video. In third experiment, our goal is to associate object trajectories across cameras to solve the multiple camera tracking problem in presence of camera motion.

4.1. Action recognition

In order to perform action recognition, we use complete set of actions given in Fig. 7 and compute the matching score given in Eq. (18). We match each action against all other actions. Action recognition is then achieved by selecting the action with the highest matching score.

In order to analyze the performance of the proposed approach against the standard approach based on the static fundamental matrix, we generated confusion matrices for both methods. In Fig. 8, we show the confusion matrices computed for our approach (part a) and the standard approach (part b), where light color illustrates similar actions and dark illustrates dissimilar actions. For automatically clustering similar actions, we use a modified version of the normalized cut segmentation method proposed by Shi and Malik [33]. In particular, we represent each action as a node in a graph with links to all other actions. The weight of the links are provided from the values stored in the confusion matrices. Clustering is then achieved by evaluating the eigenvectors of a generalized eigensystem [33]. A qualitative analysis of the clustered confusion matrices shows that using the proposed method the actions are correctly clustered. In contrast, using the standard method no clusters are obtained except for the forehand tennis stroke action where the exemplars are captured by stationary cameras. These results confirm that the proposed method based on the TFM is better for clustering the similar actions.

4.2. Action retrieval

The task in this experiment is to retrieve the occurrences of an exemplar action in a long video. The search is performed by comparing a floating temporal window with the exemplar action, where the length of the temporal window is chosen as the length of the exemplar action. In order to test the performance of the method, we used a long tennis sequence in which a tennis player is performing various

Fig. 7. The set of actions used to test the proposed method. For each action, we display the first image of the sequence along with superimposed trajectories of the landmark points.

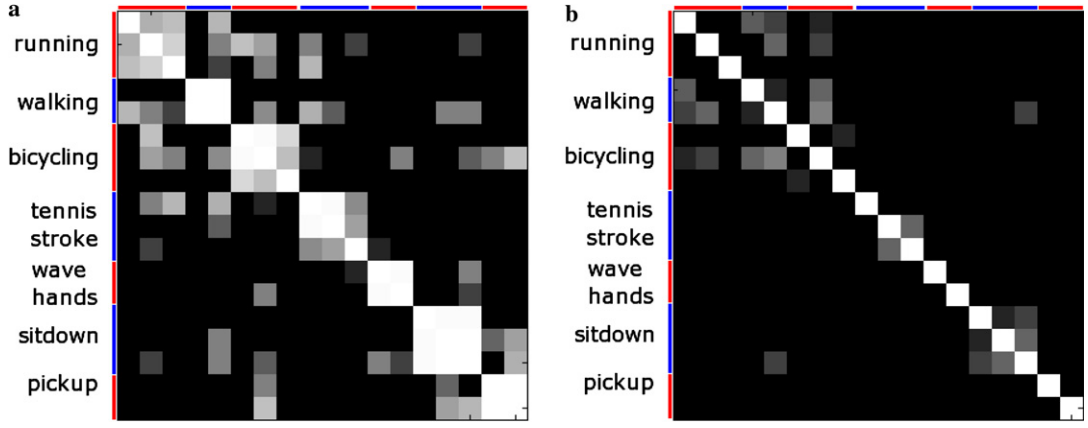


Fig. 8. Confusion matrices computed for two different methods. The light color indicates high action similarity and the dark color indicates low similarity. (a) Proposed recognition method using the matching score computed using the temporal fundamental matrix. (b) Standard epipolar geometry using the same metric. Note that correct clustering cannot be performed except for actions captured by the stationary camera.

actions, such as forehand stroke, walking, etc. In the experiment, we provide two exemplar actions: one for walking, one for tennis stroke, and retrieve them in the video sequence. As shown in Figs. 9a and b, both exemplar actions belong to two different actors and are captured by moving cameras from completely different viewpoints. In Fig. 9c, we provide both quantitative and qualitative evaluation of the method, by plotting the matching score (Eq. (18)) as a function of the frame number. On the top of the plot, the colored bars show manually extracted ground truth along with the duration of the action. The color codes of the plots correspond to the matching scores for the walking action (blue plot) and the tennis stroke (red plot). In both plots, peaks denote temporal position of the

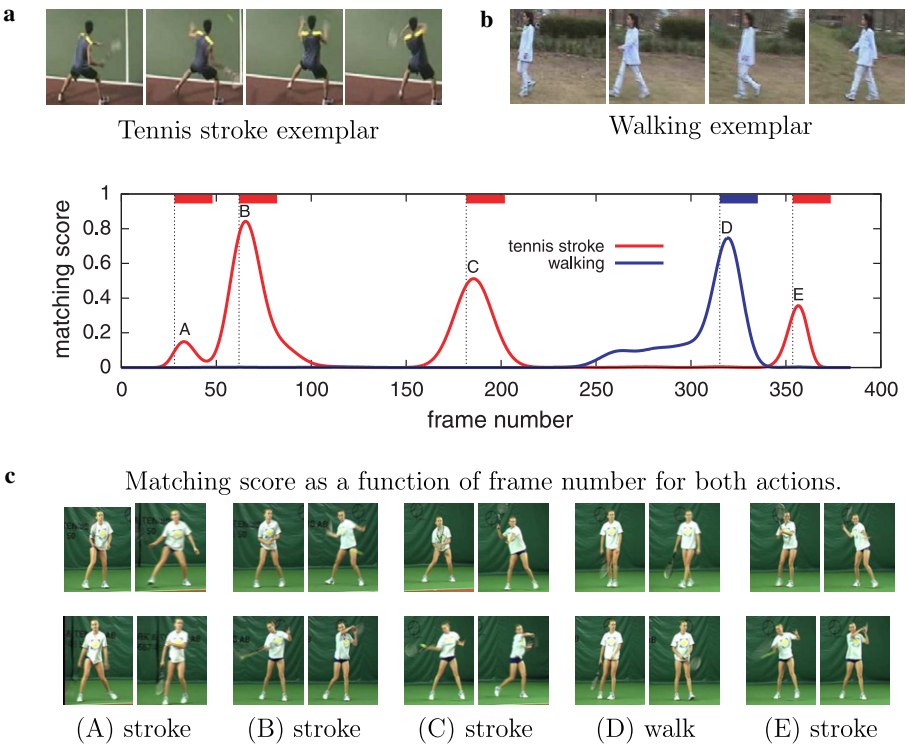


Fig. 9. Quantitative and qualitative evaluation of the action retrieval in a 400 frame long tennis sequence. In parts (a) and (b), we show the exemplar walking and tennis stroke actions. (c) Matching scores for walking (blue plot) and tennis stroke (red plot) plotted as a function of frame number. The peaks on both plots show the match of the exemplar action. We show the ground truth denoting the start and duration of the action on the top of the plot. In parts (A–E) a set of representative frames for each retrieved action are shown. The labels of each retrieved action coincide with the labels in part (c). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

action. In parts (A–E), we show sample frames of the retrieved actions (peaks in the plots). All occurrences of the exemplar actions are correctly detected.

4.3. Associating objects across cameras

In this experiment we show an application of temporal fundamental matrix in solving object association problem across multiple moving cameras. Previous approaches to associate objects across multiple cameras require scene planarity, specific camera motion or camera calibration [34–38]. These constraints limit the applicability of the proposed methods to specific environments. For instance, scene planarity can only be observed for orthographic views and is only suitable for aerial videos. Similarly, methods requiring specific camera motion such as rotation, pan, tilt, cannot work for cameras mounted on vehicles. Camera calibration is difficult to achieve and is not always available. To the best of our knowledge, there is no work on the object association problem for uncalibrated freely moving cameras.

We first define some variables. Let the set of trajectories, Γ on the left and the right cameras be denoted by $\mathbf{U}_l = (\Gamma_l^{1\top}, \ldots, \Gamma_l^{K_l\top})$, and $\mathbf{U}_r = (\Gamma_r^{1\top}, \ldots, \Gamma_r^{K_r\top})$, as given in Eq. (1). Our goal is to associate trajectories that belong to the same object in both views. Two objects are associated if and only if corresponding trajectories satisfy: $\Gamma_r^{i\top}\mathcal{F}(t)\Gamma_r^j = 0$. Under this geometric constraint, we hypothesize possible associations and compute the confidence of the hypothesized match based on the matching score given in Eq. (18). Given a set of matching hypothesis, and their confidences, we pose the object association problem as a graph theoretic problem. Let the nodes V of the graph $\mathbf{G}$ be the object trajectories in both left and right cameras views. The edges E connecting the nodes are only defined between the object trajectories of the left and the right cameras, i.e., there are no edges between the nodes corresponding to the trajectories in the same camera view (Fig. 10). Thus, the trajectory association graph is a bipartite graph. The weight $w_{i,j}$ between trajectories Γ_l^i and Γ_r^j is as given in Eq. (18) resulting in the weight matrix:

$$W = \begin{bmatrix} \mathcal{C}_{1,1} & \mathcal{C}_{1,2} & \ldots & \mathcal{C}_{1,K_l} \\ \mathcal{C}_{2,1} & \mathcal{C}_{2,2} & \ldots & \mathcal{C}_{2,K_l} \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{C}_{K_r,1} & \mathcal{C}_{K_r,2} & \ldots & \mathcal{C}_{K_r,K_l} \end{bmatrix}. \tag{19}$$

The correspondence between trajectories is achieved by computing the maximum matching of the weighted bipartite graph. A matching of a graph is a set of edges, such that no two edges share a common vertex. A maximum matching contains the highest number of possible edges. In our case, the maximum matching will provide 1-1

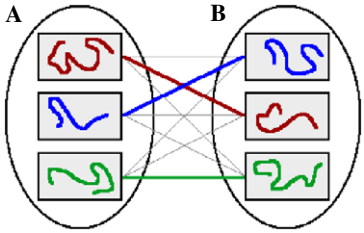


Fig. 10. *Maximum matching of weighted bipartite graphs*: trajectory correspondence is established by finding the maximum matching between two parties A and B of the graph corresponding to trajectories in the left and the right cameras. The correct trajectory correspondences are emphasized with thicker lines.

(one-to-one) mappings from A to B (see Fig. 10), such that $\sum_i \sum_j w_{i,j}$ is maximized. An important advantage of the proposed method over recovering individual fundamental matrices per frame is the requirement of fewer number of tracked objects. In this scenario, at least eight objects per frame are required to compute the standard fundamental matrix. In contrast, for the proposed method, since all observations over time are used during estimation of TFM, as minimum as "two objects" are adequate to find associations. Contrast this to at least eight point correspondence required for the standard epipolar geometry.

We have tested the proposed method on a number of sequences. In the figures, trajectories of the associated objects are shown with the same color in both views. In all the experiments, we set the degree of TFM to three resulting in thirty-six unknowns. In order to employ TFM, we use object trajectories in each camera for the first 60 to 100 frames. After associating these objects, we maintain the object labels without computing TFM again.

In Fig. 11a, we demonstrate that the objects are correctly associated for an indoor sequence where two moving hand-held cameras are viewing the lobby of a building. Next sequence, which is shown in Fig. 11b, is captured outdoor using two hand-held camcorders with a small but jerky motion. The objects are associated using the first 60 frames and labels are maintained for the remaining 300 frames. Further, we applied our method to aerial videos which were captured by two UAVs (Fig. 11c). Associations between three moving objects are correctly established. In Fig. 11d, we tested the proposed method on sequences captured by two hand-held camcorders positioned at two opposite ends of the building. One camera is almost stationary (see red trajectory on the right view) whereas the other camera translates. In parts (e) and (f), we show the recovered epipolar lines using the TFM and object associations in both camera views for the qualitative judgment. In the figures, both the points on the objects and the corresponding epipolar lines are labeled with the same color.

5. Conclusions

We proposed a novel approach for matching two actions in presence of camera motion. Proposed method makes use the epipolar geometry, however, instead of using the standard fundamental matrix, which is not able to deal with independent camera motion, we presented a new

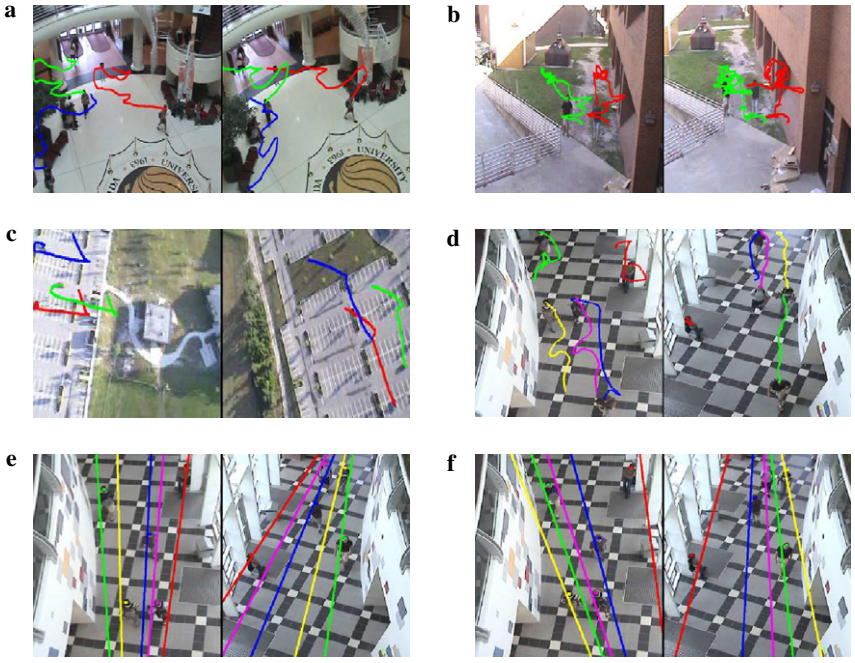


Fig. 11. (a–d) Associating objects are indicated by color codes across cameras. In all the sequences, the cameras view the scene from an oblique angle. (e–f) Epipolar lines for sequences shown in (d) at different time instants computed using the TFM. Note that the cameras are looking at the opposite ends of the scene, thats why the order of the lines is switched in two views, e.g., red object is right most in the left image but it is left most in the right image. However, the associations are correct. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

construct called ''temporal fundamental matrix.'' Each element of the temporal fundamental matrix is a function of time, which can be approximated by a polynomial assuming smooth camera motion. In order to demonstrate the versatility of the proposed approach, we presented results on action retrieval and recognition and association of objects across moving cameras.

We would like to conclude by commenting on two important issues. First, in this work, we assumed tracking problem has been already solved and we are given a set of joint trajectories of an actor performing the action. Body and joint tracking in unconstrained environment is still a complex problem, which needs to be addressed. However, recently some progress has been made (e.g., [39,40]). Second, we have assumed smooth camera motion, which is valid for short video clips of two to three seconds, which is typical length of most human actions. However, if the motion is not smooth and if the video sequence much longer temporal fundamental matrix may not be valid. In that case, one needs to compute different temporal fundamental matrix for each smooth section of video, which is still better than using a separate fundamental matrix for each frame. Also, one can experiment with higher degree of polynomials to deal with some non-smooth camera motion.

References

[1] A. Efros, A. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: ICCV, 2003.
[2] R. Polana, R. Nelson, Recognizing activities, in: ICPR, 1994.
[3] L. Zelnik-Manor, M. Irani, Event-based analysis of video, in: CVPR, 2001.
[4] I. Laptev, T. Lindeberg. Space-time interest points, in: ICCV, 2003.
[5] C. Harris, M. Stephens, A combined corner and edge detector, in: 4th Alvey Vision Conference, 1988, pp. 147–151.
[6] M. Yang, N. Ahuja, M. Tabb, Extraction of 2d motion trajectories and its application to hand gesture recognition, PAMI 24 (8) (2002) 1061–1074.
[7] M. Black, Y. Yacoob, Recognizing facial expressions in image sequences using local parameterized models of image motion, IJCV 25 (1) (1997) 23–48.
[8] A. Bobick, J. Davis, The representation and recognition of action using temporal templates, PAMI 23 (3) (2001) 257–267.
[9] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: ICCV, 2005.
[10] G. Cheung, S. Baker, T. Kanade, Articulated shape-from-silhouette and its use for human body kinematics estimation and motion capture, in: CVPR, 2003.
[11] P. Tresadern, I. Reid, Uncalibrated and unsynchronized human motion capture: a stereo factorization approach, in: CVPR, 2004.
[12] D. Weinland, R. Ronfard, E. Boyer. Motion history volumes for free viewpoint action recognition, in: IEEE International Workshop on Modeling People and Human Interaction, 2005.
[13] A. Gritai, Y. Sheikh, M. Shah, On the invariant analysis of human actions, in: ICPR, 2004.
[14] A. Yilmaz, M. Shah, Recognizing human actions in videos acquired by uncalibrated moving cameras, in: ICCV, 2005.
[15] T. Syeda-Mahmood, A. Vasilescu, S. Sethi, Recognizing action events from multiple viewpoints, in: IEEE Workshop on Detection and Recognition of Events in Video, 2001.
[16] A. Yilmaz, M. Shah, Action sketch: a novel action representation, in: CVPR, 2005.
[17] C. Rao, A. Gritai, M. Shah, T. Syeda-Mahmood, View-invariant alignment and matching of video sequences, in: ICCV, 2003.
[18] V. Parameswaran, R. Chellappa. Quasi-invariants for human action representation and recognition, in: ICPR, vol. 1, 2002, pp. 307–310.
[19] L. Wolf, A. Shashua, On projection matrices pk–p2, and their applications in computer vision, IJCV 48 (1) (2002) 53–67.
[20] S. Avidan, A. Shashua, Threading fundamental matrices, PAMI 23 (1) 73–77.
[21] A. Fitzgibbon, A. Zisserman. Multibody structure and motion: 3-d reconstruction of independently moving objects, in: ECCV, 2000.
[22] G. Johansson, Visual perception of biological motion and a model for its analysis, Percept. Psychophys. 73 (2) (1973) 201–211.
[23] R. Harris, R. Bradley, P. Jenner, M. Lappe, Neuronal Processing of Optic Flow, Elsevier, Amsterdam, 1999.
[24] J. Rehg, T. Kanade, Model-based tracking of self-occluding articulated objects, in: ICCV, 1995, pp. 612–617.
[25] V. Pavlovic, J. Rehg, T. Cham, K. Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models, in: ICCV, 1999.
[26] J.K. Aggarwal, Q. Cai, Human motion analysis: A review, CVIU 73 (3) (1999) 428–440.
[27] D.M. Gavrila, The visual analysis of human movement: a survey, CVIU 73 (1) (1999) 82–98.
[28] T. Moeslund, E. Granum, A survey of computer vision-based human motion capture, CVIU 81 (3) (2001) 231–268.
[29] C. Rao, A. Yilmaz, M. Shah, View invariant representation and recognition of actions, IJCV 50 (2) (2002) 203–226.
[30] H. Longuet-Higgins, A computer algorithm for reconstructing a scene from two projections, Nature 293 (1981) 133–135.
[31] O. Faugeras, What can be seen in three dimensions with an uncalibrated stereo rig, in: ECCV, 1992, pp. 563–578.
[32] R. Hartley, In defence of the 8-point algorithm, in: CVPR, 1995, pp. 1064–1070.
[33] J. Shi, J. Malik, Normalized cuts and image segmentation, PAMI 22 (8) (2000) 888–905.
[34] Y. Sheikh, M. Shah, Object tracking across multiple independently moving cameras, in: ICCV, 2005.
[35] J. Kang, I. Cohen, G. Medioni, Tracking objects from multiple stationary and moving cameras, Proc. of IEEE Intelligent Dist. Surveillance Systems (2004).
[36] A. Rahimi, B. Dunagan, T. Darrel, Simultaneous calibration and tracking with a network of non-overlapping sensors, in: CVPR, 2004.
[37] O. Javed, Z. Rasheed, K. Shafique, M. Shah, Tracking across multiple cameras with disjoint views, in: ICCV, 2003.
[38] R. Collins, A. Lipton, H. Fujiyoshi, T. Kanade, Algorithms for cooperative multisensor surveillance, Proceedings of IEEE 89 (10) (2001) 1456–1477.
[39] D. Ramanan, D. Forsyth, A. Zisserman, Strike a pose: tracking people by finding stylized poses, in: CVPR, 2005.
[40] A. Gritai, M. Shah, Tracking of human body joints using anthropometry, in: Int. Conf. of Multimedia and Expo, 2006.