Int J Comput Vis (2009) 84: 325–343
DOI 10.1007/s11263-009-0239-8

Matching Trajectories of Anatomical Landmarks Under Viewpoint, Anthropometric and Temporal Transforms

Alexei Gritai · Yaser Sheikh · Cen Rao · Mubarak Shah

Received: 2 January 2008 / Accepted: 15 April 2009 / Published online: 24 April 2009

Abstract An approach is presented to match imaged trajectories of anatomical landmarks (e.g. hands, shoulders and feet) using semantic correspondences between human bodies. These correspondences are used to provide geometric constraints for matching actions observed from different viewpoints and performed at different rates by actors of differing anthropometric proportions. The fact that the human body has approximate anthropometric proportion allows innovative use of the machinery of epipolar geometry to provide constraints for analyzing actions performed by people of different sizes, while ensuring that changes in viewpoint do not affect matching. In addition, for linear time warps, a novel measure, constructed only from image measurements of the locations of anatomical landmarks across time, is proposed to ensure that similar actions performed at different rates are accurately matched as well. An additional feature of this new measure is that two actions from cameras moving at constant (and possibly different) velocities can also be matched. Finally, we describe how dynamic time warping can be used in conjunction with the proposed measure to match actions in the presence of nonlinear time warps. We demonstrate the versatility of our algorithm in a number of challenging sequences and applications, and report quantitative evaluation of the matching approach presented.

Keywords Applications · Trajectory matching · Human Information Processing · Motion

A. Gritai (✉)
Cernium Corporation, Reston, USA
e-mail: agritai@cernium.com

Y. Sheikh
Carnegie Melon University, Pittsburgh, USA
e-mail: yaser@cs.cmu.edu
url: http://www.cs.cmu.edu/~yaser

C. Rao
PVI Virtual Media Services, New York, USA
e-mail: cen.rao@gmail.com
url: http://www.pvi.tv

M. Shah
School of Electrical Engineering and Computer Science,
University of Central Florida, Orlando, USA

1 Introduction

In his landmark treatise titled *Human Actions*, Ludwig Von Mises (1966) opens his first chapter with the statement, "*Human action is purposeful behavior*". He states that actions ostensibly reflect the actor's intention, conscious or unconscious. It is not surprising, therefore, that visual perception of actions is a critical cognitive function for interpreting the intention of an observed actor and for understanding the observer's environment. As social entities, the ability to interpret and predict the behavior of others is fundamental to normal human functioning. In fact, there is a growing body of evidence that humans might actually understand the actions of another individual in terms of the same neural code that they use to produce the same action themselves (Decety and Grezes 1999). It is for these reasons that the analysis of human actions is a subject of interest in a number of scientific communities such as philosophy (Goldman 1970), developmental psychology (Prinz 1997), economics (Von Mises 1966) and recently in cognitive neuroscience (Fogassi et al. 1996; Blakemore and Decety 2004). It is also why developing algorithms for action recognition must figure prominently in the pursuit of both machine intelligence and robotics.

Developing algorithms to recognize human actions has proven to be a significant challenge since it is a problem that combines the uncertainty associated with computational vi-

sion with the added whimsy of human behavior. Even without these two sources of variability, the human body has no less than 244 degrees of freedom (Zatsiorsky 2002) and modeling the dynamics of an object with such non-rigidity is not an easy task. Further compounding the problem, recent research into anthropology has revealed that body dynamics are far more complicated than was earlier thought, affected by age, ethnicity, class, family tradition, gender, skill, circumstance and choice (Farnell 1999). Human actions are not merely functions of joint angles and anatomical landmark positions, but bring with them traces of the psychology, the society and culture of the actor. Thus, the sheer range and complexity of human actions makes developing action recognition algorithms a daunting task. To develop computer algorithms for analyzing actions, it is important to identify properties that are expected to vary according to a set of transformations with each observation of an action, but which should not affect recognition:

Viewpoint. Except in specific application, it is unreasonable, in general, to assume that the viewpoint from which actions are observed would remain constant across different observations of that action. Thus, it is important that algorithms for action recognition exhibit stability in recognition despite large changes in viewpoint. The relationship of action recognition to object recognition was observed by Rao and Shah (2001), and developed further by Parameswaran and Chellappa (2002, 2003), Gritai et al. (2004) and by Yilmaz and Shah (2005). In these papers, the importance of view invariant recognition has been stressed, highlighting the fact that, as in object recognition (Verfaillie 1992), the vantage point of the camera should not affect recognition. The projective and affine geometry of multiple views is well-understood, see Hartley and Zisserman (2000), and various invariants have been proposed. There has also been some discussion of viewpoint variance and invariance in cognitive neuroscience in the context of both object and action recognition (Verfaillie 1992; Daems and Verfaillie 1999). In the proposed approach, accurate matching in the presence of varying viewpoint is a central problem which we address by using geometric relationships between the two observed executions of an action.

Anthropometry. In general, an action can be executed, irrespective of the size or gender of the actor. It is therefore important that action recognition be unaffected by "anthropometric transformations". Unfortunately, since anthropometric transformations do not obey any known laws, formally characterizing invariants is impossible. However, empirical studies have shown that these transformations are not *arbitrary* (see Kroemer et al. 1982). The study of human proportions has a great tradition in science, from the 'Golden Sections' of ancient China, India, Egypt and Greece down to renaissance thinkers like Leonardo Da Vinci (the Vitruvian Man) and Albrecht Durer, with modern day applications in Ergonomics and human performance engineering. We provide a functional definition of anthropometric transforms making implicit use of the 'laws' governing human body proportions to provide geometric constraints for matching. Instead of using a single point representation, we explore the use of several points on the actor for action recognition, and use geometric constraints with respect to two *actors* performing the action instead of two camera *views*. This innovative use of geometry allows two interesting results for the recognition of actions. The first result provides a constraint to measure the dissimilarity of the posture of two actors viewed in two images. The second result extends this first constraint to globally measure dissimilarity between two actions.

Execution Rate. With rare exceptions such as synchronized dancing or army drills, actions are rarely executed at a precise rate. Furthermore, the cause of temporal variability can be two fold, caused by the actor or possibly by differing camera frame-rates. It is desirable, therefore, that action recognition algorithms remain unaffected by some set of temporal transformations. Definition of this set of temporal transformations is dependent on application. In this paper, we propose an approach that assumes that only linear temporal transformations in time can occur, and performs detection under this assumption. Under this assumption we also propose a new metric that can match a model action to a test action despite constant velocity motion of the camera. However, in some applications of action recognition the assumption of temporal linear transformation might not be acceptable. To handle general non-linear temporal transformations we utilize Dynamic Time Warping (DTW) for matching, which ensure only that the temporal order is preserved.

In this paper, we decouple of the problem of tracking anatomical landmarks in images and the problem of matching the trajectories generated by a tracking algorithm. Given the trajectories of different anatomical landmarks of actors on a query (or test) action and a model (or pattern) action, we present a novel dissimilarity measure that determines whether the trajectories in the query video match the model action, allowing for three sets of transformations: viewpoint transformations, anthropometric transformations and temporal transformations. Figure 1 shows trajectories from the same action as captured under these different transformations. The algorithm, designed for illustration purpose, makes use of a measure that we demonstrate, both theoretically and empirically, to be able to match actions despite changes in the viewpoint of the actors to the camera. This measure is computed by looking at the eigenvalues of a matrix constructed from image measurements of anatomical landmarks. We propose a functional definition of the

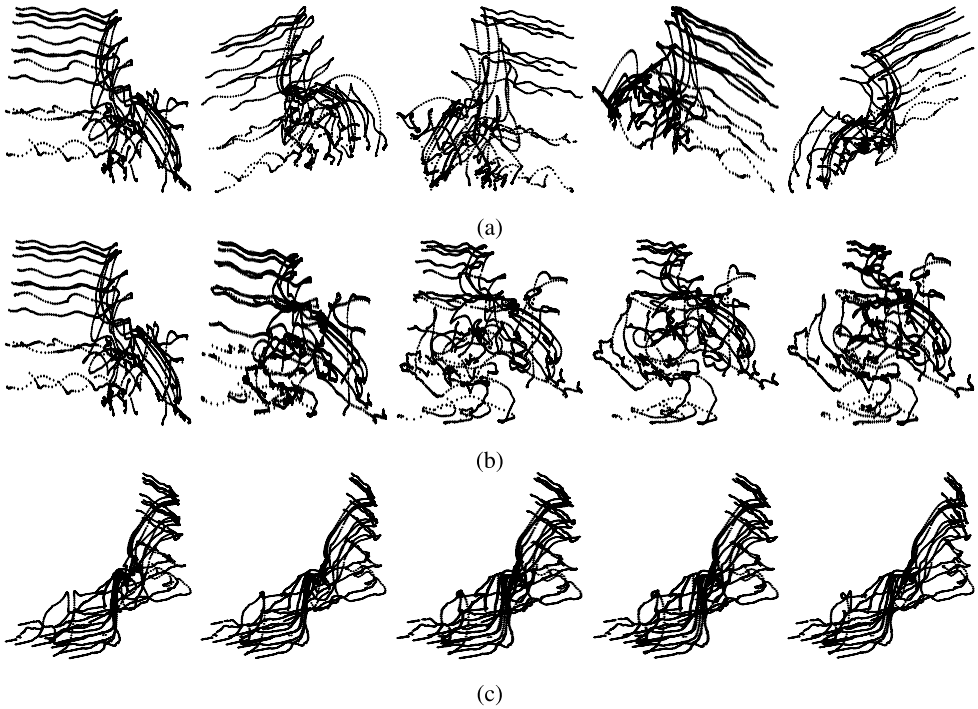

Fig. 1 Trajectories of anatomical landmarks of the same action under different types of transformation. The *first row* (**a**) presents the trajectories under different viewpoint transformations, the *second row* (**b**) under anthropometric transformations, and the *third row* (**c**) demonstrates the same trajectories obtained with different camera velocities along the X-direction

class of anthropometric transformations and use this definition to demonstrate that the measure defined is also stable to changes in the anthropometry of the actors involved. This is also demonstrated empirically during experimentation. We then use the proposed measure with DTW to determine whether two actions are the same, except by a linear transformations of time. The assumption of a stationary camera is then lifted, by allowing the cameras to move with constant velocity. We define a novel measure to match actions in this scenario which inherits all the properties of the earlier measure (stability to changes in viewpoint and anthropometry and use for temporally invariant matching). We demonstrate the application of the proposed approach in many diverse scenarios such as action synchronization, action recognition and gait analysis. Using motion capture data we also quantitatively analyze the proposed measure, verifying the properties described in the paper.

The rest of the paper is organized as follows. We situate our work in context of previous research in Sect. 2 and describe our representation and notation in Sect. 3. We then unfold the three different layers of analysis successively for viewpoint transformations (Sect. 4), for anthropometric transformations (Sect. 5) and finally temporal transformations (Sect. 6). Results are presented in Sect. 7, followed by conclusions in Sect. 8.

2 Literature Review

Research on human action recognition through computer vision started in the late seventies, the earliest work probably being the PhD thesis of Herman (1979). This work used a static representation, a stick figure in a single image, to analyze different postures of a person. The importance of dynamics was almost immediately realized and used in a series of papers in the early eighties, (O'Rourke and Badler 1980; Akita 1984; Rashid 1980). Since then a large body of literature has accumulated studying different approaches to track, reconstruct and recognize human motion. Surveys of the area have been regularly published including Liao et al. (1994), Cedras and Shah (1995), Ju et al. (1996), Aggarwal and Cai (1999) and Gavrila (1999), Moeslund and Granum (2001), Buxton (2003), and Hu et al. (2003), and finally Aggarwal and Park (2004). Under Gavrila's taxonomy of human motion analysis, methods can be roughly classified as image-based approaches or 3D approaches, i.e. methods that perform recognition directly from image measurements and those that try to recover and then analyze 3D information of human postures and dynamics. Typically in 3D approaches, models of human body and human motion are used and a projection of the model in a particular posture is then compared with each frame of the input video to recognize the action. The advantage of these approaches is that since a 3D model is explicitly used these methods are inherently view invariant. However, they are usually computationally quite expensive, (Hogg 1984) and 3D recovery of the articulated objects is still a difficult problem. As a result, 3D approaches are therefore usually limited in some specific applications, such as athletic analysis and sign language recognition (Campbell et al. 1996; Davis and Shah 1994).

In image-based approaches only 2D measurements, such as optical flow, spatio-temporal gradients or point trajectories, are computed across a sequence of frames to recognize actions. An overwhelming majority of recent work in action recognition falls in this category. The methods proposed in this category can be further subdivided into two categories: (1) Feature based approaches and (2) 'Direct' approaches.

A whole slew of different features have been proposed and used. To recognize the temporal textures, the statistical features of optical flow such as mean flow magnitude, standard deviation, the positive and negative curl and divergence, are used in Polana and Nelson (1994). Other features to recognize human activities include region-based (Davis and Bobick 1997; Niyogi and Adelson 1994; Polana and Nelson 1994; Ayers and Shah 1998; Li and Greenspan 2005), temporal trajectory based (Nishikawa et al. 1998; Yang and Ahuja 1998; Rao and Shah 2001; Gould and Shah 1989), part-based (Black and Yacoob 1995; Bregler et al. 2000; Ju et al. 1996) or a combination of these (Black and Jepson 1998; Haritaoglu et al. 2000). The approaches work based on features capturing either 2D shape or motion information. Usually, the recognition system involves some dissimilarity or similarity measurement between the activities and the models, such as the shape of the silhouettes, the trajectories of the moving hands, the point clouds from the body parts. Hidden Markov models have also been a popular tool in using these features for recognition following its success in speech recognition (Yang et al. 1997). The earliest papers included work by Starner and Pentland (1996), and Yamato et al. (1995). More sophisticated models, such as Coupled Hidden Markov Model (CHHM) (Oliver et al. 1999), Variable Length Markov Model (VLMM) (Johnson et al. 2001), Layered Hidden Markov Model (LHMM) (Oliver et al. 2002), stochastic context free grammar (SCFG) (Bobick and Ivanov 1998), and Hierarchical Hidden Markov model (HHMM) (Venkatesh et al. 2005; Singer et al. 1998), have been proposed for efficiently representing and recognizing activities from one or more persons. However these method require training data, and generally lack the capability of explaining the actions semantically.

Most recently, approaches loosely applying the paradigm of 'direct' methods proposed by Horn and Weldon (1988) which utilize the spatio-temporal information directly for motion analysis, have started to appear. The difference from feature based approaches is that image measurables are *directly* used for recognition. An approach based on the statistical features of spatio-temporal gradient direction is used for classifying human activities, e.g. walking, running, and jumping (Caspi and Irani 2000). In Zelnik-Manor and Irani (2001), an action recognition system is proposed by matching the histogram of the optical flow generated by different actions. This approach is extended in Shechtman and Irani (2005), so that the spatio-temporal volumes of actions are exploited, and a correlation measure is computed for recognizing the same action from different video. The spatio-temporal information of actions is further used for detecting irregularities in images and in video (Boiman and Irani 2005). In this work, a statistical framework is proposed for matching the patches containing actions in the video. Sukthankar et al. (2005) proposed using boosted classifiers to detect action events from the video from simple spatiotemporal filters. In Blank et al. (2005) the silhouettes of the moving subjects are used in addition to the spatio-temporal information of the pixels. The method utilizes properties of the solution to the Poisson equation to extract space-time features such as local space-time saliency, action dynamics, shape structure and orientation, furthermore, these features are used for action recognition, detection and clustering.

The fundamental drawback of using such 2D image-based approaches, direct approaches in particular, is that they are viewpoint dependent. An intermediate category of approaches, including this paper, use image measurements, but exploit 3D constraint by exploiting the geometry of multiple views. Seitz and Dyer (1997) used view-invariant measurement to find the repeating pose of walking people and the reoccurrence of position of turning points. Laptev et al. (2005) proposed using spatio-temporal points from the video to compute the fundamental matrix/homography, which are in temporal matrix format, and to detect the periodic motion once the transformation between video clips are obtained. Parameswaran and Chellappa (2009) proposed to use the 2D view invariant values, namely the cross ratio values, as the measure for matching the human actions from different viewing directions. The multiple trajectories from the joints of a person are recorded, the pose during the action is matched with a canonical body pose, and the matching coefficients are used for representing the action, and the temporal variance of the actions is compensated using DTW. Finally, the actions are matched by comparing the coefficients of the actions.

3 Notation

In this section we discuss our representation of actions and propose a novel matching scheme based on semantic correspondences between humans. Geometric constraints on these correspondences are used to analyze actions as they occur. The main concern in our work is the recognition of human activity performed by different people at varying rates in different environments or viewpoints.

3.1 Representation of Actors and Actions

The model of a moving body as a point is ubiquitous in the computer vision community. In our work, the input is

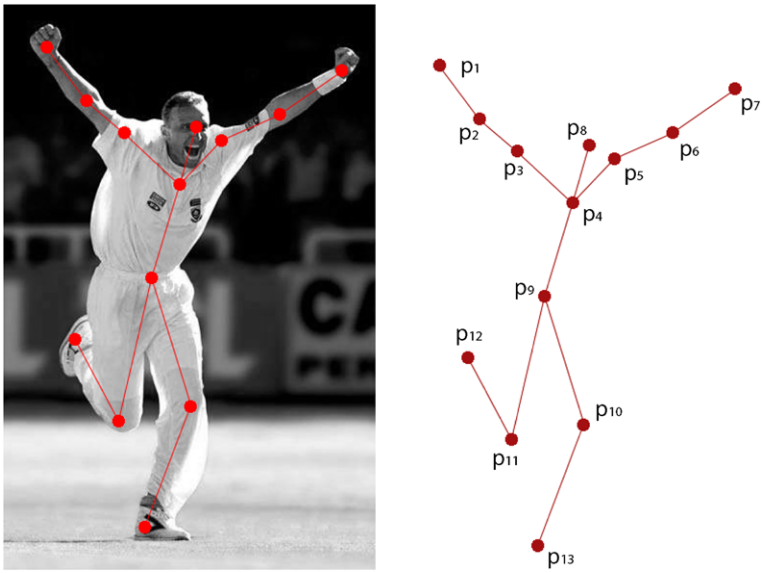

Fig. 2 Point-based representation. Experiments in Johansson (1993) demonstrate that point-based representations contains sufficient information for action recognition, and figure illustrates the landmark positions in these experiments

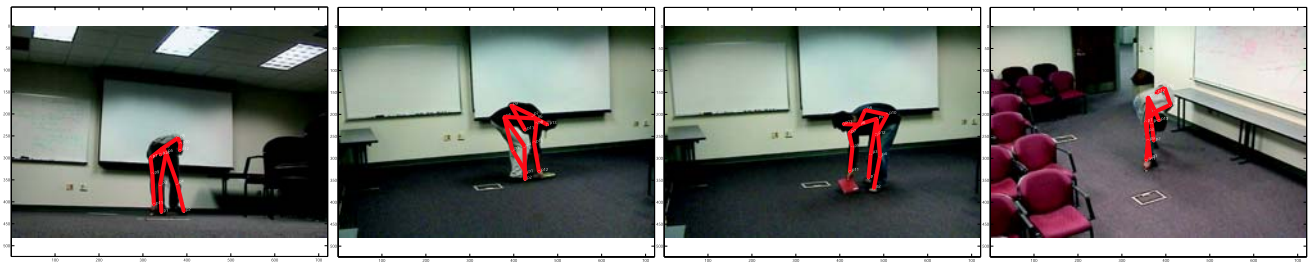

Fig. 3 Frames corresponding to 'picking up' in four sequences. The left-most frame corresponds to the model sequence, and the rest correspond to the test sequences. In each sequence, the actors are in markedly different orientations with respect to the camera, but in the same posture

the 2D motion of a set of 13 anatomical landmarks, $\mathcal{L} = \{1, 2, \ldots, 13\}$, as viewed from a camera, see Fig. 2. Johansson (1993) demonstrated that a simple point-based model of the human body contained sufficient information for the recognition of actions. Relying on this result, we represent the current pose and posture of an actor in terms of a set of points in 3D-space $\hat{\mathbf{X}} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$, where $\mathbf{X}_i = (X_i, Y_i, Z_i, \Lambda)^\top$ are homogenous coordinates and $n \in \mathcal{L}$. A *posture* is a stance that an actor has at a certain time instant, not to be confused with the actor's *pose*, which refers to position and orientation (in a rigid sense). Each point represents the spatial coordinate of an anatomical landmark (see Bridger 1982) on the human body as shown in Fig. 2. For the ith frame of the kth camera, the imaged pose and posture are represented by $\hat{\mathbf{U}}^k = \{\hat{\mathbf{u}}_1^k, \hat{\mathbf{u}}_2^k \ldots \hat{\mathbf{u}}_m^k\}$, where $\hat{\mathbf{u}}_i^k = \{\mathbf{u}_{(i,1)}^k, \mathbf{u}_{(i,2)}^k \ldots \mathbf{u}_{(i,n)}^k\}$, $\mathbf{u}_{(i,j)}^k = (u_{(i,j)}, v_{(i,j)}, \lambda)^\top$ and m is the number of frames. $\hat{\mathbf{X}}$ and $\mathbf{u}^k$ are related by a 4×3 projection matrix $\mathbb{C}^k$, i.e. $\mathbf{u}^k = \mathbb{C}^k \hat{\mathbf{X}}$. As will be seen presently, nine imaged points on human body are required in each frame of video and, at least, one of them must correspond to the body part directly involving in action. We refer to each entity involved in an *action* as an *actor*. An *action element*, $\hat{\mathbf{u}}_t$, is the portion of an action that is performed in the interval between frames t and $t + 1$. Each action is represented as the set of action elements. For a comparison of other representations to this one the reader is referred to Gavrila (1999).

4 Viewpoint Transformations

Figure 3 shows the same action ('picking up a book') from 4 different points of view. Although the same action is being performed the distribution of points on the image differs significantly. As has been observed previously for object recognition, it is usually unreasonable to place restrictions on the possible viewpoint of the camera, and action recognition algorithms should therefore demonstrate *invariance* to changes in viewpoint. Invariants are properties of geometric configurations that are unaffected under a certain class of transformations. It is known that view-invariants do not exist for general 3D point sets (Burns et al. 1992). However, there are useful properties that are not strictly invariant, but remain stable over most transformations. We now describe a

measure to match actions that is based on one such property. Assuming two frames are temporally aligned (until Sect. 6), the labels associated with each anatomical landmark provide point-to-point correspondence between the two postures. The constraint we use is that if the two imaged point sets match they are projections of the same structure in 3D. In Rao and Shah (2001), a rank constraint based dissimilarity measure was described that was stable to camera viewpoint changes. The main drawback of this dissimilarity measure was the assumption of affine cameras. To remove this assumption, instead of using this factorization based rank constraint, we use a constraint derived from epipolar geometry. For the projective camera model, the fundamental matrix (a 3×3 matrix of rank 2), $\mathbf{F}$, is defined between corresponding points by

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix}^T \mathbf{F} \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = 0, \tag{1}$$

for the pair of matching points $(u, v) \leftrightarrow (u', v')$ in trajectories, observed from two different viewpoints. Clearly, given a fundamental matrix, we can use (1) to measure the dissimilarity between two trajectories, so that the squared residual for all points is minimized. By rearranging (1) a dissimilarity measure can also be defined directly from the trajectory values themselves (without explicitly computing $\mathbf{F}$). Given at least 9 point matches, we have,

$$\mathcal{A}\mathbf{f} = \begin{bmatrix} u'_1u_1 & u'_1v_1 & u'_1 & v'_1u_1 & v'_1v_1 & u'_1 & u_1 & v_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u'_tu_t & u'_tv_t & u'_t & v'_tu_t & v'_tv_t & u'_t & u_t & v_t & 1 \end{bmatrix} \mathbf{f} = 0, \tag{2}$$

where $\mathbf{f} = [\, f_{11}\ f_{12}\ f_{13}\ f_{21}\ f_{22}\ f_{23}\ f_{31}\ f_{32}\ f_{33}\,]^T$ is the fundamental matrix vectorized in row-major order. We refer to $\mathcal{A}$ as the observation matrix, which is constructed using only the coordinates of points of corresponding 2D trajectories. Since (2) is homogenous, for a solution of $\mathbf{f}$ to exist, matrix $\mathcal{A}$ must have rank at most eight, and this fact can be exploited to measure dissimilarity. Of course, due to the noise or the matching error, the rank of matrix $\mathcal{A}$ may not be exactly eight. The condition number of $\mathcal{A}$, i.e. the ratio of the smallest singular value, σ_9, to the largest singular value, σ_1, of $\mathcal{A}$ provides the algebraic error of corresponding points in matrix $\mathcal{A}$. This ratio can be used to measure the match of two trajectories,

$$\kappa = \frac{\sigma_9}{\sigma_1}. \tag{3}$$

It should be noted that the observation matrix $\mathcal{A}$, and therefore this dissimilarity metric, is constructed only from measured image position. In addition to viewpoint changes caused by different camera locations, anthropometric transformations are also expected, caused by different actors, which is discussed next.

5 Anthropometric Transformations

Both body size and proportion vary greatly between different races and age groups and between both sexes. However, while human dimensional variability is substantial, several anthropometric studies (see Easterby et al. 1982; Bridger 1995; Badler et al. 1993) empirically demonstrate that it is not *arbitrary*. These studies have tabulated various percentiles of the dimensions of several human anatomical landmarks. In this paper, we conjecture that for a large majority of the human population the proportion between human body parts coupled with a rigid transformation in 3D space can be captured by a projective transformation of $\mathbb{P}^3$, projective 3-space (Hartley and Zisserman 2000).

Conjecture 1 *Suppose the set of points describing actor* A_1 *is* $\hat{\mathbf{X}}$ *and the set of points describing actor* A_2 *is* $\hat{\mathbf{Y}}$. *The relationship between these two sets can be described by a matrix* $\mathcal{M}$ *such that*

$$\mathbf{X}_i = \mathcal{M}\mathbf{Y}_i, \tag{4}$$

where $i = 1, 2 \ldots n$ *and* $\mathcal{M}$ *is a* 4×4 *non-singular matrix.*

This was empirically supported using the quite representative data in Bridger (1982, Table 5-1 and 5-2) which record the body dimensions of male and female workers between the ages of 18 and 45). For the most extreme case, between the dimensions of the '5th percentile woman' and the '95th percentile man', where a mean error of 227.37 mm was found before transformation, a mean error of 23.87 mm was found after applying an appropriate transformation. Using this property, geometric constraints can be used between the imaged points, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ of the two actors. The transformation $\mathcal{M}$ simultaneously captures the different pose of each actor (with respect to a world coordinate frame) as well as the difference in size/proportions of the two actors.

5.1 Postural Constraint

If two actors are performing the same action, the postures of each actor at a corresponding time instant with respect to the action time coordinate should be similar. Thus an action can be recognized by measuring the dissimilarity of posture at each corresponding time instant.

Proposition 1 *If $\hat{\mathbf{x}}_t$ and $\hat{\mathbf{y}}_t$ describe the imaged posture of two actors at time t, a matrix $\mathcal{F}$ can be uniquely associated with $(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)$ if the two actors are in the same posture.*

It is known (p. 247 Sect. 9.2, Hartley and Zisserman 2000) that for uncalibrated cameras the ambiguity of structure is expressed by such an arbitrary non-singular projective matrix. If two actors are in the same posture, the only difference between their point-sets is a projective relationship (Conjecture 1). Thus, if an invertible matrix $\mathcal{P}$ exists between $\mathbf{X}$ and $\mathbf{Y}$, i.e. $\mathbf{Y} = \mathcal{P}\mathbf{X}$, a *fundamental* matrix is uniquely determined by $\mathbf{x}^\top \mathcal{F}\mathbf{y} = 0$ (Theorem 9.1, Hartley and Zisserman 2000).[1] It is important to note that the matrix $\mathcal{F}$ does not capture only the relative positions of the cameras as does the fundamental matrix $\mathbf{F}$, but instead the relative poses of the actors and the relative anthropometric transformation between the actors.

Since the labels of each point are assumed known, *semantic* correspondences (i.e. the left shoulder of A_1 corresponds to the left shoulder of A_2) between the set of points are also known. Proposition 1 states that the matrix computed using these semantic correspondences between actors inherently captures the difference in anthropometric dimensions and the difference in pose. This point is illustrated in Fig. 4. The matrix $\mathcal{F}$, computed between the actors, captured an anatomical relationship between the actors as well as the different views of the actors. The result is that the dissimilarity measure, described in Sect. 4, remains stable despite changes in anthropometry of the actors. Since the anthropometric proportions of actor can be expected to remain the same over short periods of time this fact can be used to provide an even stronger constraint which we now describe.

5.2 Action Constraint

Along with the frame-wise measurement of postural dissimilarity, it is observed here that a strong *global* constraint can be imposed on the point sets describing two actors if they are performing the same action.

[1] Points that lie on the line joining the principal points are excluded.

Proposition 2 *For an action-element $\hat{\mathbf{u}}_t$, the fundamental matrices associated with $(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)$ and $(\hat{\mathbf{x}}_{t+1}, \hat{\mathbf{y}}_{t+1})$ are the same if both actors perform the action element defined by $\hat{\mathbf{u}}_t$.*

Based on Conjecture 1, we can say that $\mathcal{M}$ remains the same between time t and $t+1$. In other words, $\mathcal{M}$ determines $\mathbf{Y}$ with respect to $\mathbf{X}$ and does not depend on the motion of $\mathbf{X}$. Since $\mathcal{M}$ is the same then the matrices, $\mathcal{F}_t$ and $\mathcal{F}_{t+1}$, corresponding to $(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)$ and $(\hat{\mathbf{x}}_{t+1}, \hat{\mathbf{y}}_{t+1})$ are the same (p. 235 Result 8.8, Hartley and Zisserman 2000).

What this means is that if both individuals perform the same action-element between frame f_t and frame f_{t+1}, the transformation that captured the difference in pose and dimension between the two actors remains the *same*. As a direct consequence, the subspace spanned by the measurement matrix A also remains the same and this suggests that if a measurement matrix were constructed using *all* the corresponding points over the entire action $\hat{A} = [A_1, A_2, \ldots A_k]$, $\kappa_{\hat{A}}$ can be used as a global measure of *action* dissimilarity. The second row of Fig. 4 illustrates this. Both actors of clearly different anatomical proportion perform the same action element (they moved their right foot back and raised their right hand). The matrix $\mathcal{F}$ computed between the actor in their original postures was used to compute epipolar lines after the execution of the action element. Clearly, to the extent that the same action element was performed, the geometric relationship is preserved. Thus, instead of considering the action as the successive motion of 13 points over n frames, each action is considered to be a cloud of $13n$ points, each point having a unique spatio-temporal index (see Fig. 1). However, the analysis thus far has assumed that temporal transformations had been accounted for. In practice, temporal transformations, small or large, always exist. We now describe how to compensate for these transformations during action analysis.

6 Temporal Transformations

While invariance to change in viewpoint is required in action analysis due to the imaging process, invariance to temporal transformations is needed due to the nominal uniqueness of each actor's execution of an action. In this paper, we describe matching algorithms that are stable under two types of transformations. First, we describe a new metric that can match actions despite linear transformation in time (scaling and shifts). We show that this metric can also match actions despite constant velocity motion of the camera. This model works effectively for many applications, particularly when the pattern is of a short duration. It was found that the use of a linear model is also appropriate

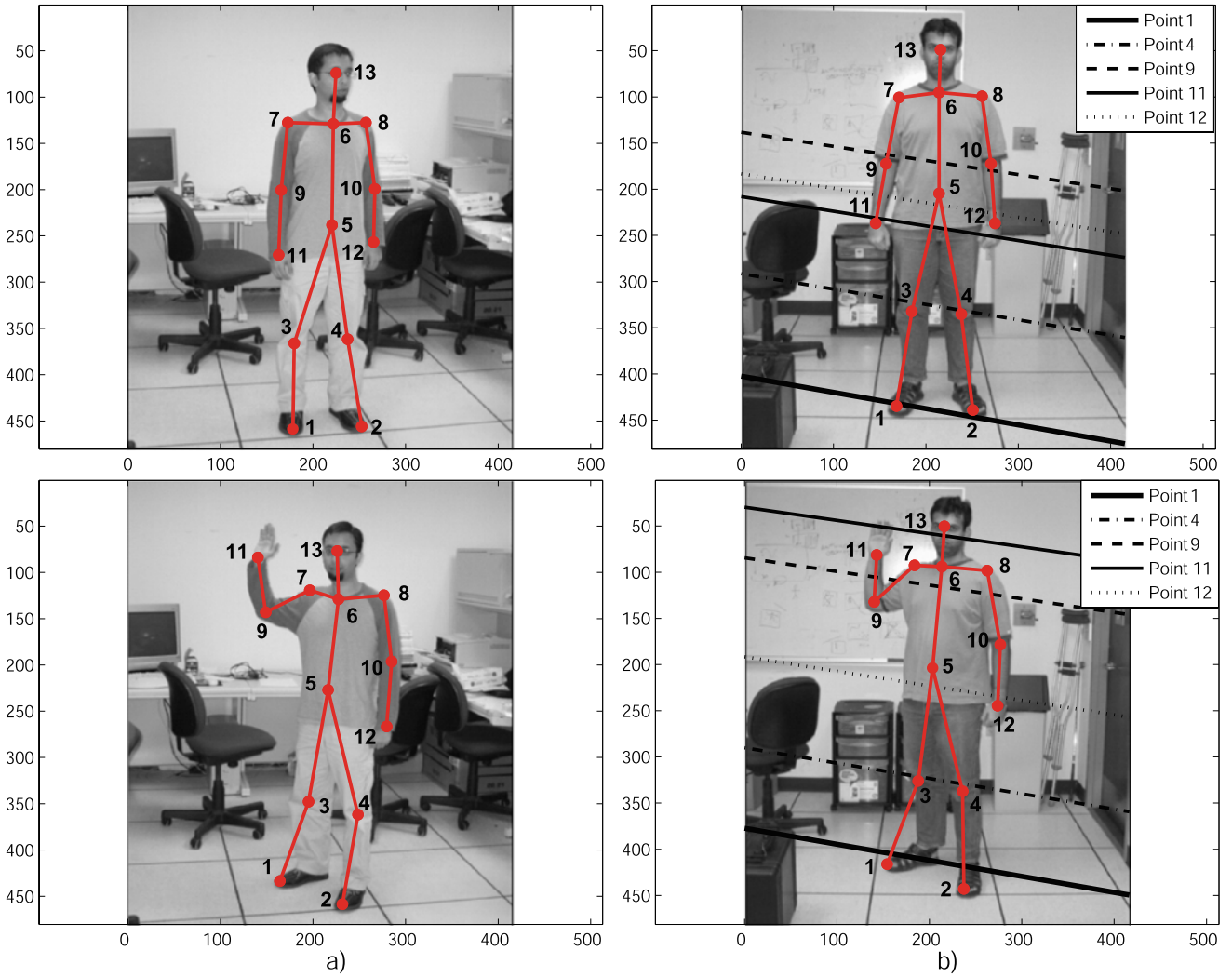

Fig. 4 The matrix $\mathcal{F}$ can capture the relationship between body landmarks of two different actors of different height, weight, etc. but in the same posture. It captures the variability in proportion as well as the change in viewpoint. (**a**) Actor 1 in two frames of the model video. (**b**) Actor 2 in the corresponding frames of the test video. The landmark correspondences in first frames of model and test video (the *first row*) were used to compute the matrix $\mathcal{F}$. The image on right in (**b**) shows epipolar lines corresponding to landmarks in the image on right in (**a**). It is clear that the landmarks in the test video lie close to the corresponding *epipolar lines*; in particular, the *epipolar lines* pass close to their corresponding landmarks 9 and 11 after the right hand of the Actor 2 was moved

for *coarse* matching and synchronization. Second, we describe a more general approach, where the temporal transformation may be highly nonlinear, using DTW to compensate for temporal transformations. In this case, there is no clearly defined class of temporal transformations, except that temporal order must be preserved during the transformation.

6.1 Linear Transformation and Constant Velocity

A linear transformation of time can be expressed as,

$$t' = a_1 t + a_2, \tag{5}$$

where a_1 is a scaling and a_2 is a shift in time. Given a model action and a test action, if we wish to deduce whether the actions observed in both sequences were equivalent up to a linear temporal transformation. In addition to differing rates of action execution, it is important to note that two cameras might have a different frame rate, and the starting points of the video in two cameras might also be shifted relatively in time. Furthermore, to remain stable despite constant velocity motion of the camera, we use the fundamental constraint of linear motion (Sheikh et al. 2007) between cameras moving independently with constant velocity. As shown in Sheikh et al. (2007), the relationship between points from the two sequences can be expressed as,

$$\mathcal{A}_{\mathbf{T}} f = \begin{bmatrix} u'_1 t_1 & u'_1 u_1 & u'_1 v_1 & u'_1 t_1 v_1 & u'_1 t_1 u_1 & u'_1 & v'_1 t_1 & v'_1 u_1 & v'_1 v_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u'_n t_n & u'_n u_n & u'_n v_n & u'_n t_n v_n & u'_n t_n u_n & u'_n & v'_n t_n & v'_n u_n & v'_n v_n \end{bmatrix}$$
$$\begin{bmatrix} v'_1 t_1 v_1 & v'_1 t_1 u_1 & v'_1 & t_1 & u_1 & v_1 & t_1 v_1 & t_1 u_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v'_n t_n v_n & v'_n t_n u_n & v'_n & t_n & u_n & v_n & t_n v_n & t_n u_n & 1 \end{bmatrix} f = 0, \tag{6}$$

where f is a 18-dimensional vector and $\mathcal{A}_{\mathbf{T}}$ is a matrix constructed from time-space image coordinates of the corresponding points. If points exactly correspond to each other, then the rank of $\mathcal{A}_{\mathbf{T}}$ is 17, otherwise, the 18th singular value will be non zero. Thus instead of estimating κ from the observation matrix associated with the original fundamental matrix, we construct this new observation matrix and use the condition number of this matrix as our measure of dissimilarity. Thus to determine if the temporal transformation between the two observations is linear, despite constant velocity motion of the camera, κ can be used.

6.2 Non-linear Transformation

Finally, we describe the use of DTW to compensate for non-linear temporal transformations. Dynamic Time Warping is a widely used method for warping two temporal signals (Sakoe and Chiba 1978). It uses an optimum time expansion/compression function to perform a non-linear time alignment. The applications include speech recognition, gesture recognition (Darrell et al. 1995), signature verification and for video alignment (Rao et al. 2003). DTW is particularly suited to action recognition, since it is expected that different actors may perform some portions of an action at different rates, relatively. The use of DTW is not trivial in this case since both the local (postural) constraint and the global (action) constraint need to be incorporated in computation of the dissimilarity measure. Applying a temporal window (k frames before and after the current one) for computation of dissimilarity measure between two agents provided a marked improvement.

To synchronize two signals I and J by DTW, a distance, $\mathbf{E}$, is computed to measure the misalignment between two temporal signals, where $\mathbf{E}(i, j)$ represents the error of aligning signals (distance measure) up to the time instants t_i and t_j respectively. The error of alignment is computed incrementally using the formula:

$$\mathbf{E}(i, j) = \mathbf{dist}(i, j) + \mathbf{e}, \tag{7}$$

where

$$\mathbf{e} = \min\{\mathbf{E}(i-1, j), \mathbf{E}(i-1, j-1), \mathbf{E}(i, j-1)\}.$$

Here $\mathbf{dist}(i, j)$ captures the cost of making time instants t_i and t_j correspond to each other. The best alignment is then found by keeping track of the elements that contribute to the minimal alignment error at each time step and backward following a path from element $\mathbf{E}(i, j)$ to $\mathbf{E}(1, 1)$.

Similar to Rao et al. (2003), in our framework I and J are trajectories representing similar or different actions observed from distinct viewpoints, and by introducing $\kappa(i, j)$ as the $\mathbf{dist}(i, j)$, the standard DTW becomes appropriate for action recognition and robust to view, anthropometric and temporal transformations.

7 Experimental Results

To demonstrate the performance of the approach in this paper, we performed experiments both qualitatively in several challenging scenarios and quantitatively using motion capture data. All data used during recognition was in the form of image measurements from uncalibrated cameras. In the qualitative experiments we demonstrate the versatility of the proposed approach in solving a variety of problems including action recognition, video synchronization and gait analysis. We designed our experiments to test each 'layer' of analysis in isolation as well as experiments that demonstrate efficacy under all sources of variability.

7.1 Qualitative Results

In this set of experiments, trajectories from an exemplar action were matched against trajectories from a longer test sequence. To match we manually marked the landmarks and computed κ at each frame number between the exemplar trajectories and an equally sized trajectory set (through temporal windowing) from the longer test sequence, centered around that frame number.

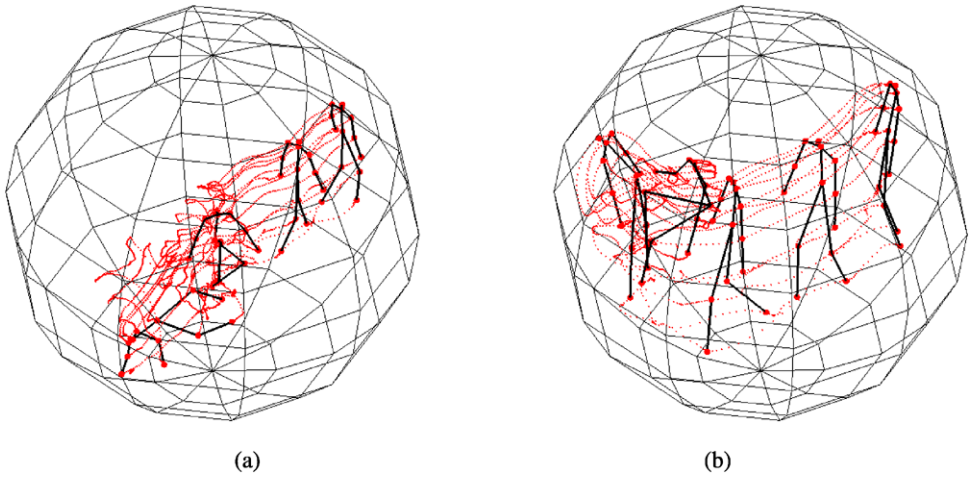

Fig. 5 Viewing spheres. (**a**) The action 'getting up' is viewed at regular intervals on a sphere around the action. (**b**) The action 'Sit Down' is viewed at regular intervals on a sphere around the action

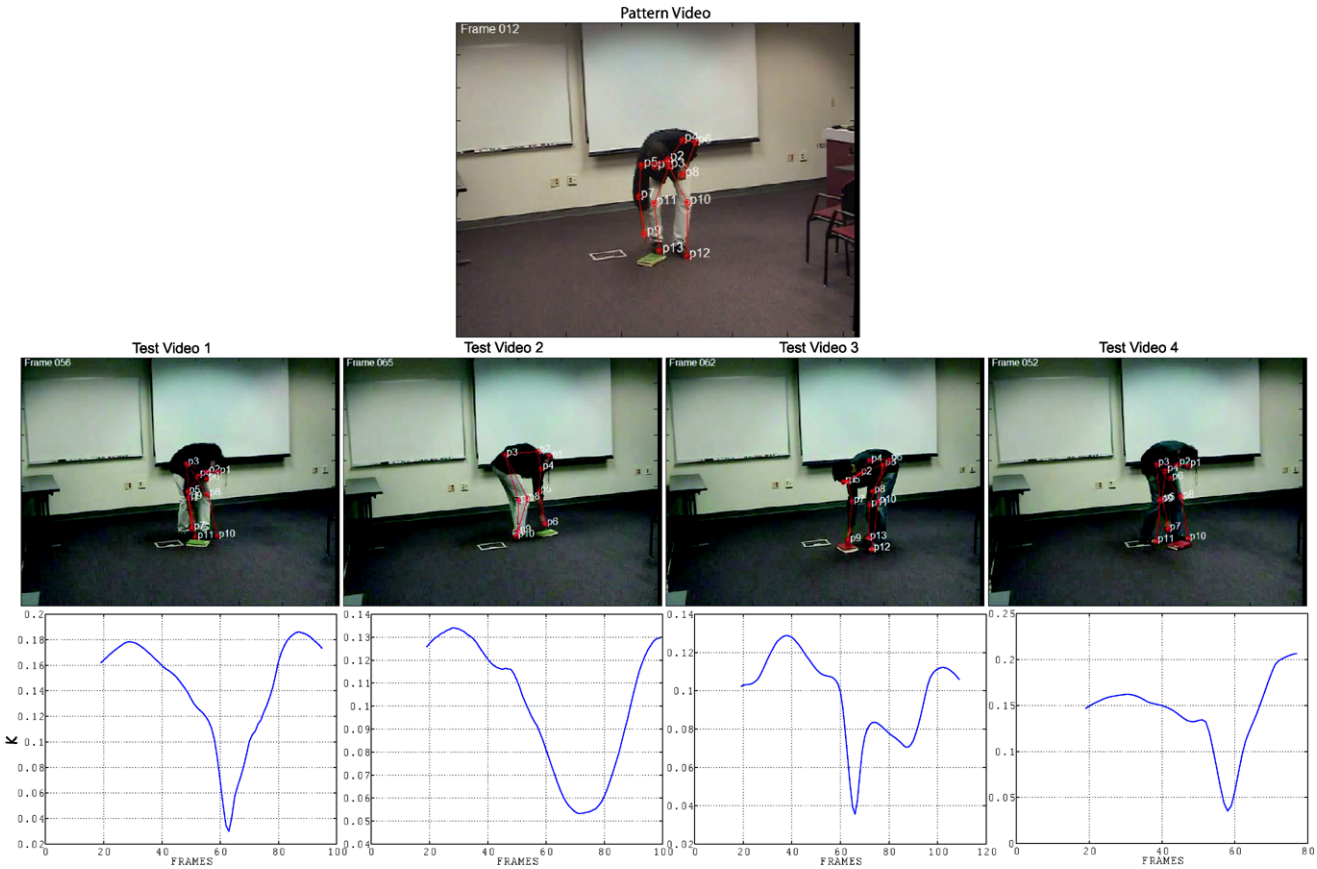

Fig. 6 Action matching from multiple views. Plots of the matching score against the frame number for four videos. Frames corresponding to the minima are shown for each video

7.1.1 Trajectory Matching

In this experiment, actors performed a sequence of three actions: walking, picking up an object, and walking away. Videos were taken of two different actors as they performed this sequence with different orientations relative to the camera. The action of picking up an object was detected in each video by matching an exemplar sequence containing only the 'picking up an object' action. Figure 6 shows plots of the matching score against frame number. The value at each time location in the plots were obtained by matching the temporal neighborhood against the exemplar sequence. It can be seen that a distinct minimum occurs at the temporal location where the best match occurs. The cor-

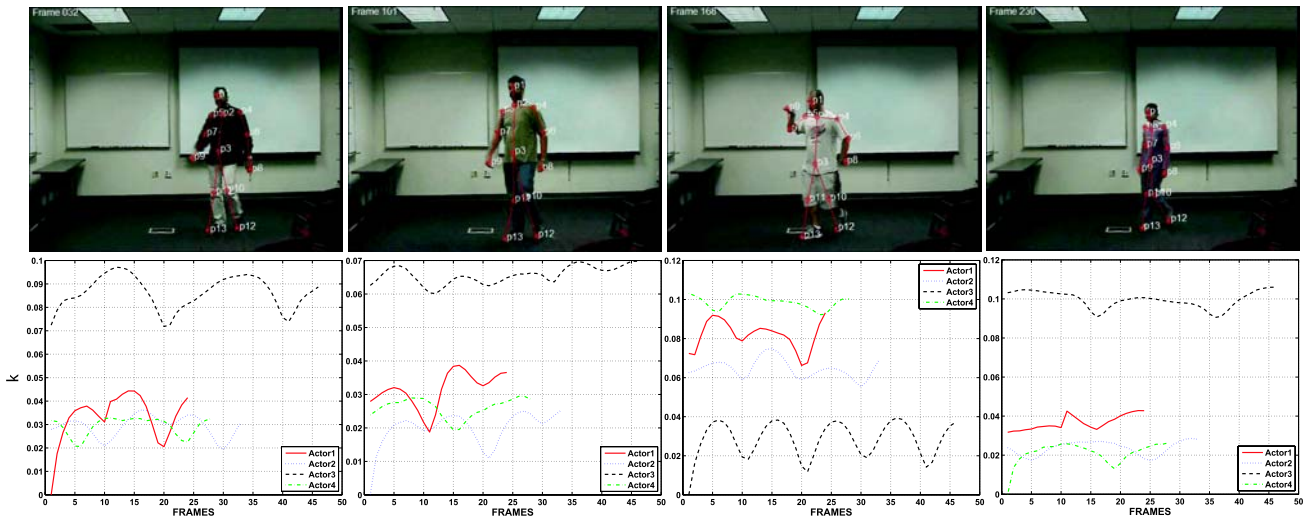

Fig. 7 Odd One Out. Actor three, the third figure from the *left*, corresponds to the actor performing the 'Egyptian' gait

Fig. 8 Following the leader (the left-most actor). The *top row* shows four frames, 22, 25, 27, and 29 before synchronization. Notice the difference in postures of each actor within each single view. The *bottom row* shows corresponding frames (to the *top row*) from the rendered sequence after synchronization

responding frames at these minima are also shown in the figure.

The sensitivity of matching was also tested in a sequence containing four individuals walking. A test pattern of a single cycle of the distinctive 'Egyptian' gait was compared to each actor's motion and the variation of the smallest singular value over time for each of the four actors is shown in Fig. 7 (the odd-one-out is the third actor from the left). There are two points of interesting in this figure. First, since the posture involved in the 'Egyptian' gait is relatively distinct from the usual human gait the smallest singular value for the third actor is consistently larger and distinct from the other actors. Second, the sinusoidal nature of the plot clearly shows the periodicity that is associated with walking. In order to generate the plots, a cycle of one gait was matched against all other gait sequences.

7.1.2 *Video Synchronization*

Three actors jumped asynchronously in the field of view of a stationary camera. The objective in this experiment was to align the actors jumps and twists so that a new *synchronized* sequence could be rendered. The temporal transformation between actors was highly nonlinear, and DTW, with a 10-frame window around the current frame and κ as the distance measure, was used. Accurate synchronization was achieved and Fig. 8 shows the result of synchronization with respect to the left-most actor using the proposed approach. The top row shows the original sequence and the bottom row

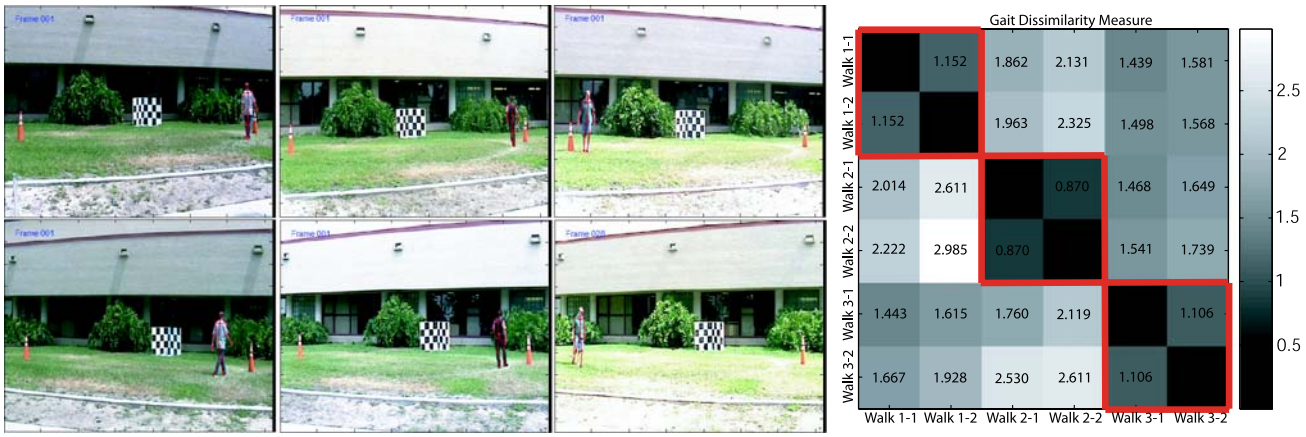

Fig. 9 Confusion Matrix for Gait Analysis. In the table the first and second columns correspond to the first actor in the first and second view respectively, and so on. The notation 1-1 refers to 'Actor 1, View 1' etc. Lower values correspond to the same actor's gaits in different views (1-1 matches best with 1-2, 2-1 with 2-2, 3-1 with 3-2)

shows the rendered sequence. The application of this sort of rendering includes post-processing of dance or exercise videos.

7.1.3 Gait Analysis

Videos of three actors were captured walking from two different viewpoints using two cameras, and, on average, each video was more than 200 frames in length. Six feature points, hands, knees and feet, were tracked. A short fragment (40 frames) was extracted from each video. The goal of this experiment was determining if the extracted fragment could be found in the video by computing the smallest singular value as the best dissimilarity measure. The table of Fig. 9 shows the confusion matrix of each gait in each view. In the table the first and second columns correspond to the first actor in the first and second view respectively, and so on. The block-diagonal nature of the confusion matrix indicates that the distance between the gait of an actor in first view and in the second view is always lower than the gait of other actors in any view.

7.2 Trajectory Matching with Real Tracking Noise

In this experiment the landmarks tracks were generated using an improved version of body joints tracking algorithm (Gritai and Shah 2006). In addition, we also manually marked the correct landmark positions to estimate the error of the joints tracking algorithm. The goal was to analyze the performance of the proposed method in a presence of the noise (detection error) with uncontrolled statistical parameters. The model action was 50 frames long and observed from the frontal view. The test actions were performed by four actors with anthropometric proportion significantly different from the actor who performed the model action. All test sequences consisted of the same set of actions, one of which was identical to the model action. The test actions 1, 2, 3 and 4 were 450, 524, 463 and 471 frames long, respectively. For each test action, all clips, starting from frame 1 and consisting of 50 consecutive frames, were compared to the model action. The first row of Fig. 10 shows the results of detecting model action in the test actions. In order to demonstrate the error in these clips, the detection error of the selected landmarks in each frame was summed up and averaged over 50 frames. The second row of Fig. 10 shows the average detection error of the landmarks selected on the left and right shoulders, elbows and wrists. The third row of Fig. 10 shows the detection error averaged over 13 detected landmarks, smallest and largest detection error in each clip among 13 landmarks. In the presence of significant error in landmark detection in each frame and successful correct action detection, we see that proposed method is indeed robust to noise introduced by real body tracking algorithms.

7.3 Quantitative Results

The following experiments quantitatively demonstrate that the proposed method is stable to changes in viewpoint, anthropometry and temporal behavior. A set of experiments were performed to evaluate each of these three properties in isolation, followed by experiments evaluating performance under all three transformations simultaneously. Motion capture data was used to provide 3D data which was projected and used in all experiments. Since the 3D coordinates of the points were known, 2D image coordinates were obtained by generating projection matrices around a viewing sphere as shown in Fig. 5. In all the experiments, actions were observed from 360 different locations in upper hemisphere, which means the elevation and azimuth were changed from 0 to 90 and from 0 to 350 degrees respectively at ten degree

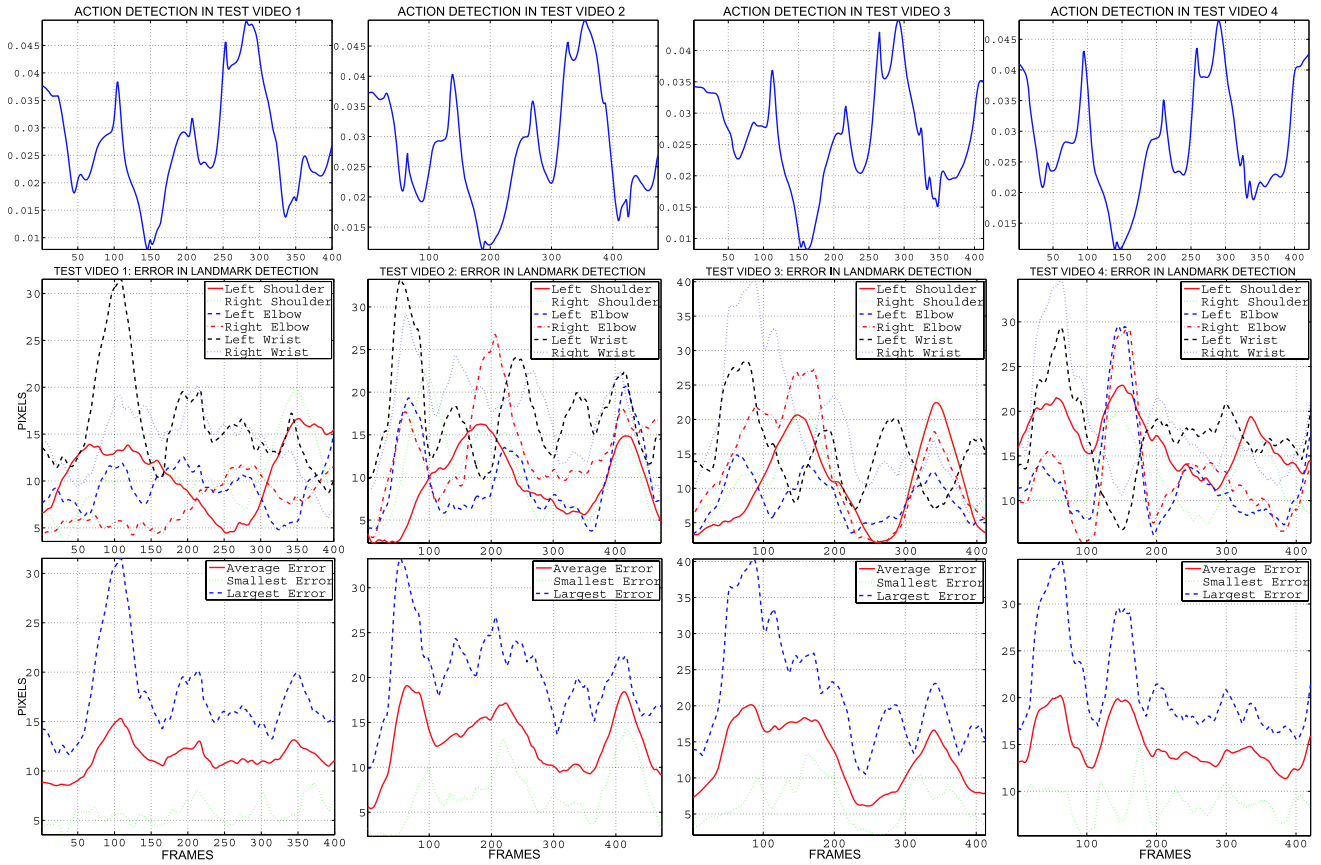

Fig. 10 The *first row* of the figure shows the results of detecting model action in the test video. The model action was detected around frame 150, 180, 160 and 150 of the test video 1, 2, 3 and 4, respectively. The *second row* shows the Euclidian error of detecting landmarks on the left and right shoulders, elbows and wrists in each frame of the test video. The *third row* shows the minimal, average and maximal Euclidian error of detecting all 13 landmarks in each frame of the test video

increments. Thus, a pair of angles, elevation and azimuth, corresponds to any of 360 possible camera locations.[2]

7.3.1 Viewpoint

In this experiment we tested the performance of the system with respect to changes in viewpoint. We demonstrated that the dissimilarity measure allows sufficient discrimination between matches and mismatches, despite different viewpoints. The first row of Fig. 1 shows the input point cloud, representing the 'getting up' action, under different view projective transformations. The experimental performance is also tested with respect to increasing noise in the measurements.

We first experimented with noiseless data obtained through motion capture equipment and rendered data at regular intervals over the described viewing sphere. To demonstrate the robustness to changes in viewpoint we recorded the log of the condition number of $\mathcal{A}_{\mathbf{T}}$ and the log of the ratio of the second smallest singular value to the largest singular value in Fig. 11(a). This figure shows that regardless of view angles the dissimilarity measure (left half of the matrix or first 360 values on the horizontal axis) is very close to zero and significantly lower than the ratio of the second smallest singular value to the largest singular value (right half of the matrix). From the illustration, one can notice that the diagonal elements are especially low. The diagonal entities correspond to the case when both camera views are exact the same.

Within this matrix, there are blocks of low values, the indices of both axes are between 325 and 360. These values correspond to the case when elevation angles of both cameras, facing to the ground, are 90 degrees, and is a special case. From our experiment, in this case the values of the two ratios are approximately 4.3×10^{-22} and 1.3×10^{-20}, while in all other matches the mean of the dissimilarity measure is 1.4×10^{-16} and the mean of the other ratio is 5.4×10^{-4}.

[2]The elevation and azimuth corresponding to some camera location n, where $n = 1, \ldots, 360$, was calculated as floor$((n-1)/36) \times 10$ and mod$((n-1)/36) \times 10$ respectively, e.g., if $n = 239$, then the elevation and azimuth are 60 and 220 degrees respectively.

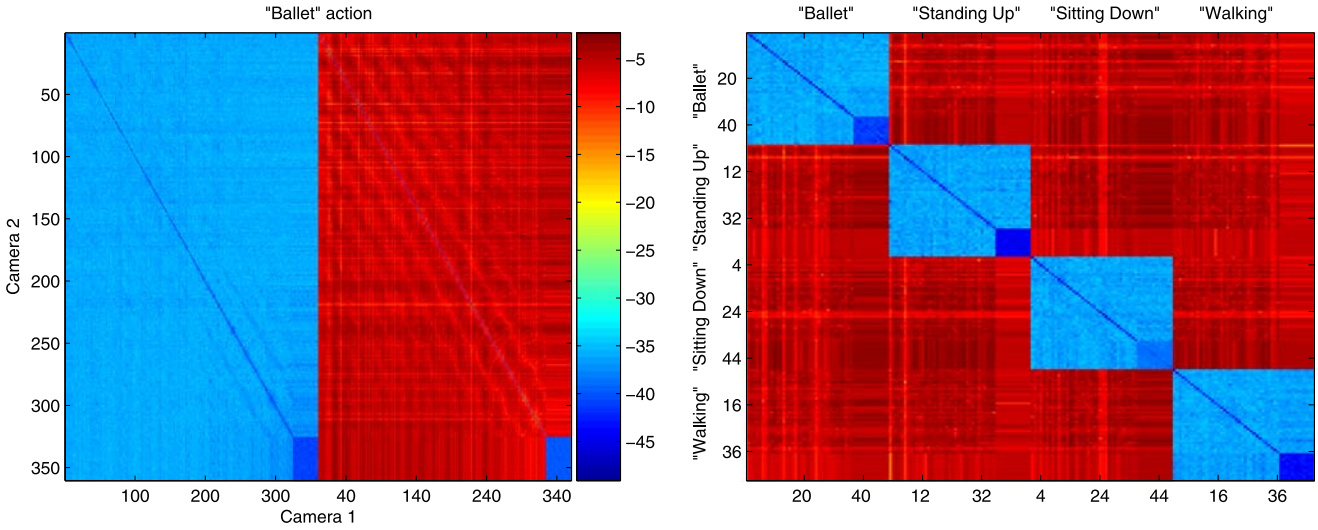

Fig. 11 Four different actions were compared to itself. The pattern (exemplar) and test actions were observed from any angle of the upper hemisphere. The left-most figure shows a significant drop between two ratios σ_9/σ_1 (*blue*) and σ_8/σ_1 (*red*), thus σ_9/σ_1 can be considered as a dissimilarity measure. There are small rectangle areas of very low values corresponding to both ratios and lying between 325 and 350 indices. It occurs when both cameras has the elevation angle of 90 degrees, which corresponds to the upper point of hemisphere. Since at the upper point of hemisphere camera centers coincide, it becomes a degenerate case. The right-most figure shows the change of σ_9/σ_1, when under different view-projective transformations, four different actions were compared to each other. The low diagonal values of the proposed dissimilarity measure demonstrate the correct discrimination among actions

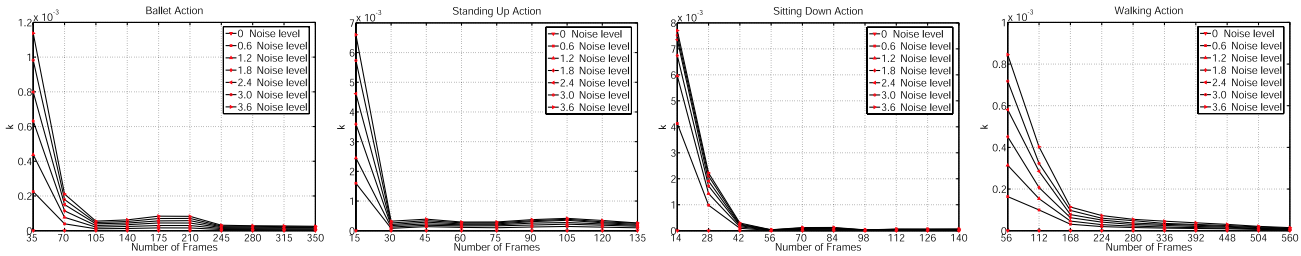

Fig. 12 The measure is robust to changes in viewpoint (different markers differentiate the different noise levels). This figure shows how the proposed dissimilarity measure changes with respect to the level of noise and the view angle. Patterns of four actions were captured at the same view point, azimuth = 30 and elevation = 25 degrees, and test actions were observed by the stationary camera at view point corresponding to the azimuth = 130 and elevation = 45 degrees. Six levels of noise, sampled from zero mean normal distribution with σ varying from 0.6 to 3.6, were added to the 2D image coordinates. Regardless of the action, when the length of the action increases, the dissimilarity measure approaches zero. The X-axis shows the number of frames, and the Y-axis shows the values of κ

Figure 11(b) shows a confusion matrix using $\log \kappa$ in a second series of experiments, where different actions were compared to each other. Four actions ('ballet', 'standing up', 'sitting down' and 'walking') were rendered from 360 different viewpoints and the block diagonal structure of the confusion matrix shows the discrimination achieved using the proposed measure. It is important to note that even in the special case mentioned above κ provides ample discrimination between different actions.

On these four actions, 'ballet', 'standing up', 'sitting down' and 'walking', we also tested the sensitivity of the metric with respect to noise and its behavior with respect to an increase in number of frames. The experimental results are presented in Fig. 12. The pattern actions were all observed from a fixed viewpoint—the azimuth and elevation were 30 and 25 degrees respectively. The test actions were observed from a significantly distinct view angle, the azimuth and elevation were 130 and 45 degrees respectively. Six levels of noise, sampled from a zero-mean normal distribution with σ varying from 0.6 to 3.6, were added to the test actions. Twenty five samples were generated from at each noise strength and the mean error at each noise level was recorded. As expected estimates of κ become more reliable as the number of frames increases, and the number of frames after which κ is stable, varies from action to action, depending largely on the 'content' of the action.

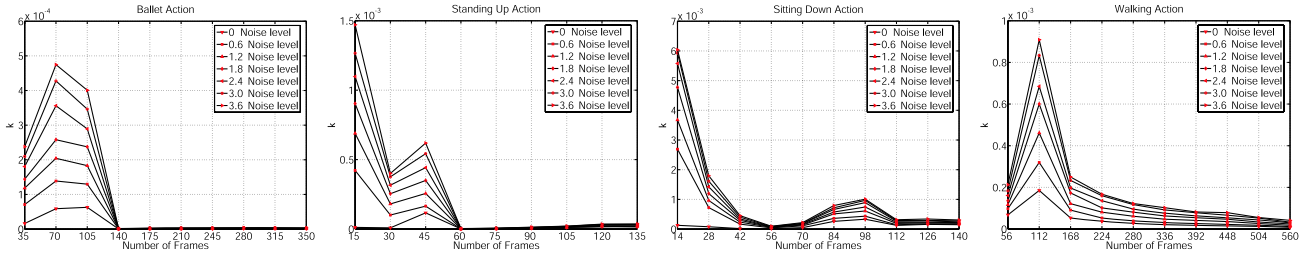

Fig. 13 Stability with respect to anthropometric transformation. The figure shows how κ changes with respect to the level of noise and the length of the action. Exemplars of four actions were captured at the same view point, azimuth = 30 and elevation = 25 degrees, and test actions were observed by the moving camera at the different view point, azimuth = 130 and elevation = 25 degrees. Six different levels of noise, sampled from a zero-mean normal distribution with σ varying from 0.6 to 3.6, were added to the 2D image coordinates. Regardless of the action, when the length of the action increases, the dissimilarity measure approaches zero

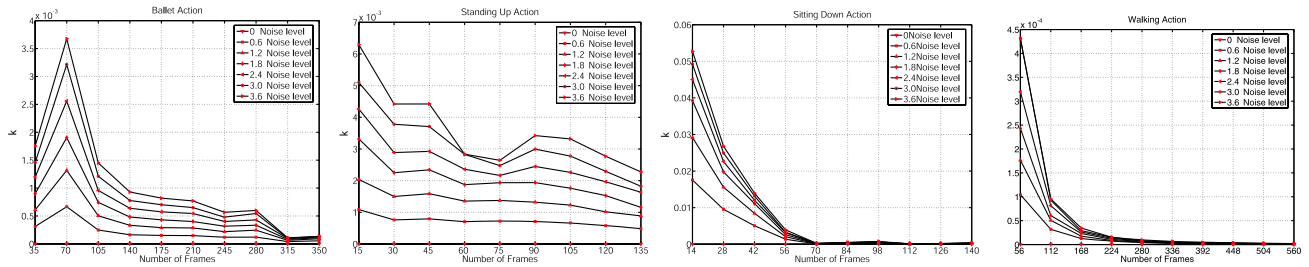

Fig. 14 The proposed dissimilarity measure is stable to temporal distortion. The figure shows how the proposed dissimilarity measure changes with respect to the level of noise and the length of the action. Patterns of four actions were captured at the same view point, azimuth = 30 and elevation = 25 degrees, and test actions were observed by the moving camera at the different view point, azimuth = 130 and elevation = 25 degrees. Six different levels of noise, sampled from a zero-mean normal distribution with σ varying from 0.6 to 3.6, were added to the 2D image coordinates. Regardless of the action, when the length of the action increases, the dissimilarity measure is approaching to zero

7.3.2 Anthropometry

In this experiment we examined the performance of κ with respect to change in the anthropometry of the actor. The second row of the Fig. 1 shows the 'getting up' action under different anthropometric transformations. Figure 13 presents the experimental results. The pattern action was observed from a view point with a fixed elevation of 60 degrees, while the azimuth was changed from 0 to 350 degrees. Similarly, the test action was observed from a view point with a fixed elevation of 30 degrees, while the azimuth was changed from 0 to 350 degrees. A 4×4 matrix $\mathcal{M}$ was randomly generated, and the whole action was transformed by $\mathcal{M}$. After 3D projective transformation, 3D points were projected onto image plane and distorted by six different levels of noise. Noise parameters were the same as in the previous set of experiments. The results showed κ to be robust to noise, and estimates of κ became more reliable as the number of frames were increased. As in the previous experiment, the number of frames after which κ stabilized, varied from action to action and depended on the action.

7.3.3 Execution Rate

This set of experiments demonstrates the robustness to temporal transformation of actions. Figure 14 shows the results. Exemplars of four actions were observed from a constant viewpoint—the azimuth and elevation were 30 and 25 degrees respectively. Test actions were observed at different view angle corresponding to the azimuth 130 and elevation 45 degrees. The test actions were distorted temporally by generating a pair (a_1, a_2) and by the same six levels of noise specified earlier. Once again we note that the longer the greater the distinctive content the action the more robust the matching.

7.3.4 Simultaneous Distortion of Temporal Index, Viewpoint and Anthropometry

The last series of experiments was performed both on rendered motion capture data and real imaged data. In these experiments we aimed to analyze the performance of κ for application in action recognition.

The first set of experiments performed on the synthetic data. The results presented in Fig. 15 demonstrate the behavior of the dissimilarity measure, κ, with respect to all

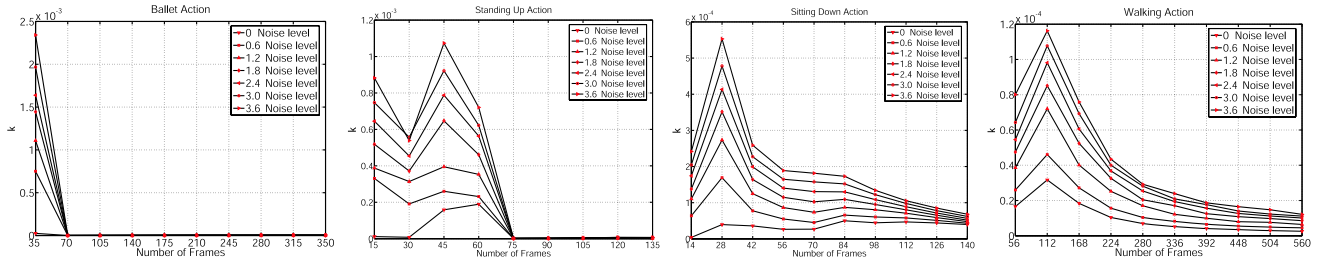

Fig. 15 The dissimilarity measure is robust to temporal, anthropometric and view distortion. This figure shows how the dissimilarity changes with respect to the level of noise and the length of the action. From *left* to *right*, four figures correspond to 'ballet', 'standing up', 'sitting down' and 'walking actions'. Patterns of four actions were captured by the moving camera with a fixed orientation, azimuth = 30 and elevation = 45 degrees, and test actions were observed by the stationary camera at the view point with azimuth = 130 and elevation = 10 degrees. Six different levels of noise, sampled from the normal distribution with means from 0.6 to 3.6 and $\sigma = 1$, were added to the 2D image coordinates. When the length of the action increases, the dissimilarity approaches zero. The X-axis shows the number of frames, and the Y-axis shows the values of κ

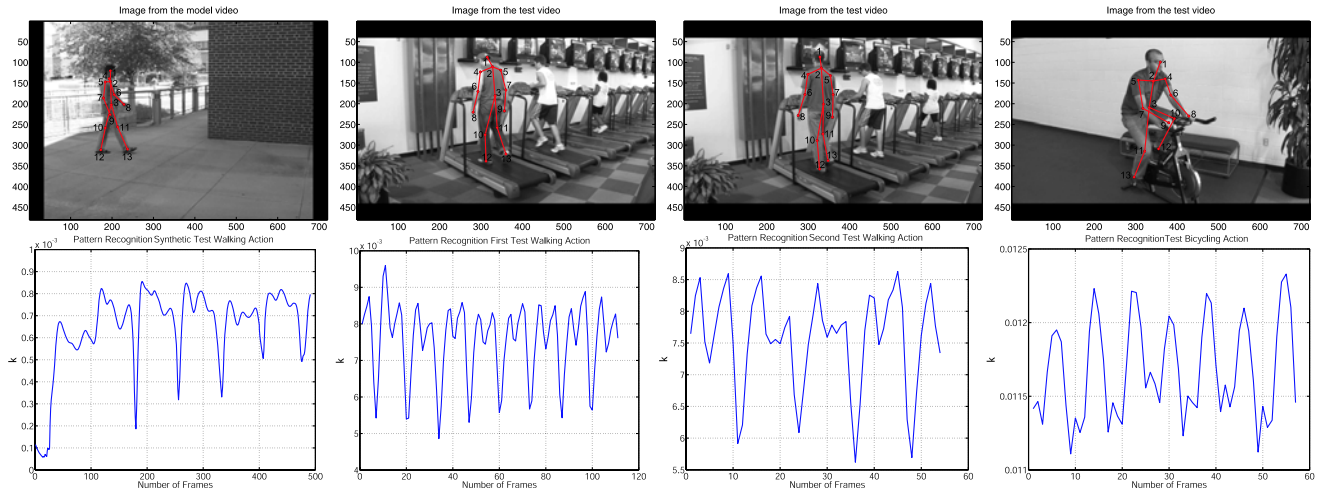

Fig. 16 The *first row* shows images from real video. The left-most image corresponds to the model action, and the remaining three correspond to the test actions. The *second row* shows results of pattern detection. The left-most figure corresponds to recognition in synthetic video. The length of the model and test video was 70 and 564 frames respectively. Two central figures shows detection of walking in real video containing walking actions. The model, and two test video were 42, 374 and 202 frames respectively. In model and test video points on bodies were marked in each third frame. The right-most figure shows the result of detection of walking action in real video that did not contain any walking actions. The test video was 212 frames long. The values of local minima in the right-most figure are greater than ones in two central figures

three types of transformations. Exemplars of the four actions ('ballet', 'standing up', 'sitting down' and 'walking') were captured by a virtual camera moving at constant velocity but a fixed orientation, azimuth = 30 and elevation = 45 degrees. Test actions were captured by the virtual stationary camera at the view point with azimuth = 130 and elevation = 10 degrees. Similar to the previous experiments, six different levels of noise sampled from the normal distribution with means from 0.6 to 3.6 and $\sigma = 1$ were added to the image coordinates. All results show robustness of κ with respect to noise.

The second set of experiments was performed on trajectories generated by the walking action. One cycle of the walking action, performed by Actor 1, was a pattern action and was captured outdoors by a stationary camera. The model action was 42 frames long. It is important to note that during the action, the pose of the actor was changing relative to the camera position. Since we consider points on the actor's body only, the stationary camera can be interpreted as a *moving* camera, and moving actor can be considered stationary. The first test clip (570 frames) was chosen from a motion capture data set. We synthesized a camera, virtually moving in 3D, and projected original data on the image plane of that camera. The second and third clips depicted Actors 2 and 3 performing "walking action" on a treadmill, and the fourth clip depicted the Actor 3 performing "bicycling action" on a recumbent bicycle. The test actions, two walking and bicycling actions, were 374, 202 and

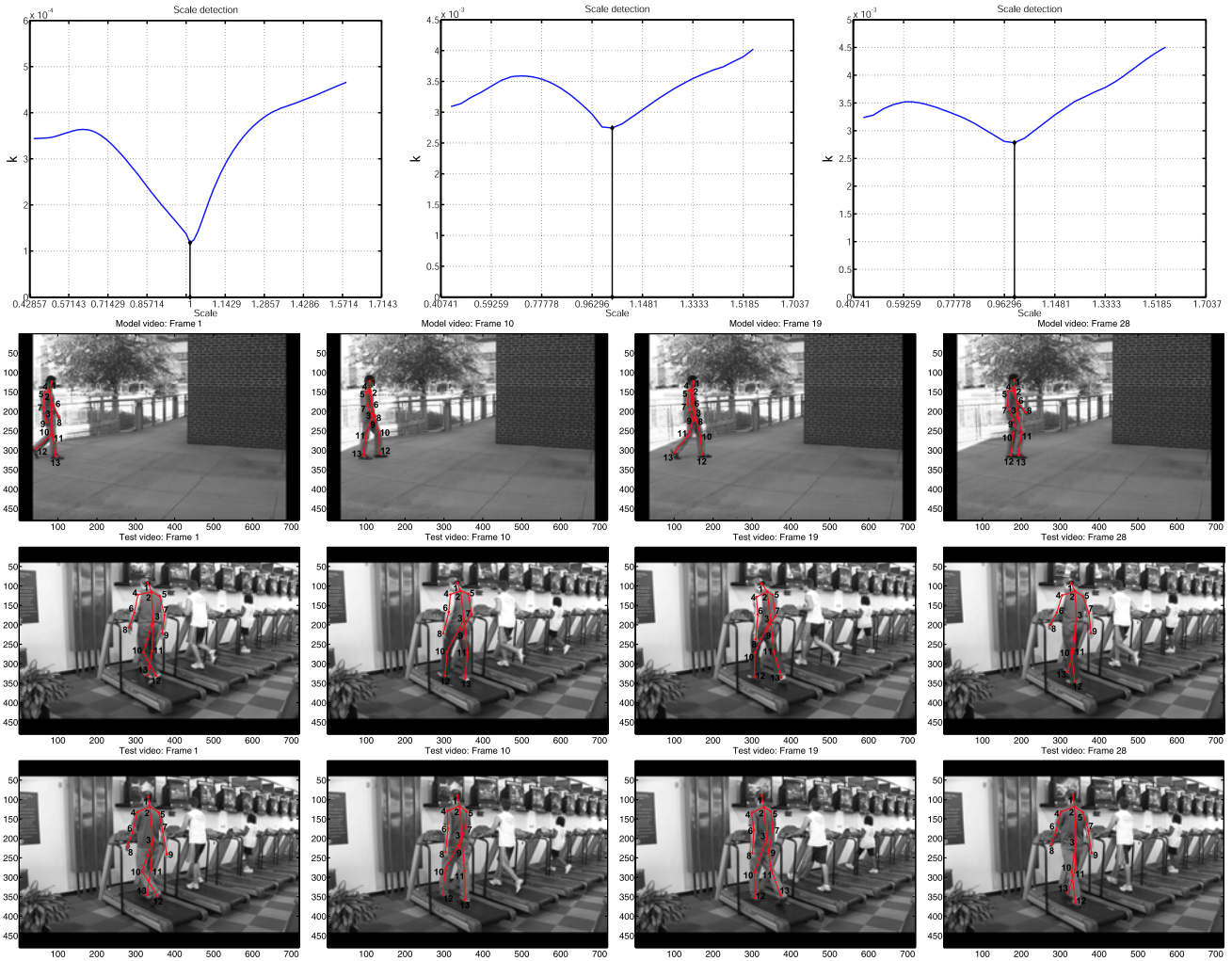

Fig. 17 The *top row* shows the results of temporal scale detection. The left-most figure shows the result of detection in synthetic video. The best match corresponds to the point where the scale is one. The remain two figures show results of scale detection in real video. Since both actors were walking faster than in the model, the best matchings correspond to the scales, which are slightly greater than one. The remaining rows show the corresponding frames after synchronization. The *second-top row* shows frames from the model video, and others show frames from the test video

212 frames, respectively. The goal of the experiments was to determine whether the query actions, captured by moving and stationary cameras, contain the pattern action, captured by stationary camera. The results are presented in Fig. 16. The leftmost image of the first row shows Actor 1 performing the pattern action. The remain three images of the first row correspond to the query actions performed by Actors 2 and 3, respectively. The second row shows the variation of κ as the pattern action was shifted in time over the duration of the test actions. The left-most figure shows the result of pattern detection in the clip from Motion capture data set, and other three figures show the results of pattern detection in second, third and fourth clips, respectively. The rightmost figure shows results of a video that did not contain any walking actions. Since that video depicted the bicycling action, we do observe some periodicity. However, the values corresponded to each potential action occurrences (local minima) were greater than values corresponded to action occurrences in video depicted walking action (see central figures).

While the above experiments determined only the location of action in test video (time translation), the final set of experiments determined the scale of temporal transformation. From all action occurrences in the previous experiment, only one occurrence was chosen in each video. Figure 17 demonstrates the results. The leftmost figure from the top row shows the result obtained on synthetic video. The best match was detected when the scale of temporal transformation was one, and this coincides with the ground truth. The other two figures from the top row show the results obtained on real video. Compared to the model action, in both test videos, actors were walking slightly faster, which was captured by the scale factor. In order to get the best match,

actions from the test video were scaled to match the model action. Analyzing results from both synthetic and real video, it is easy to see that the global minima in the left-most figure is more distinct when compared to the other two. This is attributed to noise, the length of the model action and our assumption, which is that we know the beginning point of action and do not know where action ends. As soon as a test fragment contains the action, κ becomes less sensitive to increase in scale. This effect is still observable in synthetic video but to a lesser degree. The remaining three rows show the corresponding frames after synchronization between model action and test fragments.

8 Conclusion

In this paper we have addressed the analysis of trajectories of anatomical landmarks in the presence of three key sources of distortion: viewpoint of observation, anthropometric proportion of actors, and differing rates of execution. We demonstrate, first theoretically and then empirically, that the algorithm based on the proposed dissimilarity measure is stable with respect to changes in all three distortions. During experimentation, we examine each source of distortion in isolation, followed by an evaluation in the presence of simultaneous distortion and report the quantitative performance. In addition, we provide several qualitative examples demonstrating the applicability of the proposed approach. We show various applications of proposed approach, such as video synchronization, computer aided training, and human action recognition.

References

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

Aggarwal, J., & Park, S. (2004). Human motion: modeling and recognition of actions and interactions. In *Second international symposium on 3d data processing, visualization and transmission*, 2004.

Akita, K. (1984). Image sequence analysis of real world human motion. *Pattern Recognition*, *17*(1), 73.

Ayers, D., & Shah, M. (1998). Recognizing human actions in a static room. In *Proc. IEEE workshop on applications of computer vision, WACV'98* (pp. 42–47) 1998.

Badler, N., Philips, C., & Webber, B. (1993). *Simulating humans*. London: Oxford University Press.

Black, M., & Jepson, A. (1998). Eigentracking: robust matching and tracking of articulated objects using a view-based representation. In *European conference on computer vision* (pp. 63–84) 1998.

Black, M., & Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *IEEE international conference on computer vision* (pp. 374–381) June 1995.

Blakemore, S., & Decety, J. (2004). From the perception of action to the understanding of intention. *Nature Reviews*, *2*(1), 561.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space-time shapes. In *Proceedings of the IEEE international conference on computer vision* (pp. 1395–1402) 2005.

Bobick, A. F., & Ivanov, Y. (1998). Action recognition using probabilistic parsing. In *Proceedings of the IEEE international conference on computer vision and pattern recognition* (pp. 196–202) Santa Barbara, CA, 1998.

Boiman, O., & Irani, M. (2005). Detecting irregularities in images and in video. In *Proceedings of the IEEE international conference on computer vision*, Oct. 2005.

Bregler, C., Hertzmann, A., & Biermann, H. (2000). Recovering non-rigid 3d shape from image streams. In *IEEE international conference on computer vision and pattern recognition* (pp. 13–15) 2000.

Bridger, R. (1982). *Human performance engineering: a guide for system designers*. New York: Prentice Hall.

Bridger, R. (1995). *Introduction to ergonomics*. New York: McGraw-Hill.

Burns, J., Weiss, R., & Riseman, E. (1992). The non-existence of general-case view-invariants. In J. Mundy & A. Zisserman (Eds.) *Geometric invariance in computer vision*.

Buxton, H. (2003). Learning and understanding dynamic scene activity: a review. *Image and Vision Computing*, *21*, 125–136.

Campbell, L. W., Becker, D. A., Azarbayejani, A., Bobick, A. F., & Pentland, A. (1996). Invariant features for 3d gesture recognition. In *Proceedings, international conference on automatic face and gesture recognition* (pp. 157–162) 1996.

Caspi, Y., & Irani, M. (2000). A step towards sequence-to-sequence alignment. In *Proceedings of the IEEE international conference on computer vision and pattern recognition* (pp. 682–689) 2000.

Cedras, C., & Shah, M. (1995). Motion-based recognition: a survey. *Image and Vision Computing*, *13*(2), 129–155.

Daems, A., & Verfaillie, K. (1999). Viewpoint-dependent priming effects in the perception of human actions and body postures. *Visual Cognition*, *6*, 665–693.

Darrell, T. J., Essa, I. A., & Pentland, A. P. (1995). Task-specific gesture analysis in real-time using interpolated views. *IEEE Transactions on Pattern Analysis and Machine Vision*, *18*(12), 1236.

Davis, J., & Bobick, A. (1997). The representation and recognition of action using temporal templates. In *IEEE international conference on computer vision and pattern recognition* (pp. 928–934) 1997.

Davis, J., & Shah, M. (1994). Three-dimensional gesture recognition. In *Proc. of Asilomar conference on signals, systems, and computers*, 1994.

Decety, J., & Grezes, J. (1999). Neural mechanisms subserving the perception of human actions. *Trends in Cognitive Sciences*, *3*, 172.

Easterby, R., Kroemer, K., & Chaffin, D. (1982). *Anthropometry and biomechanics—theory and application*. New York: Plenum Press.

Farnell, B. (1999). Moving bodies, acting selves. *Annual Review of Anthropology*, *28*, 341.

Fogassi, L., Gallese, V., Fadiga, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*(2), 593.

Gavrila, D. M. (1999). The visual analysis of human movement: a survey. *Computer Vision and Image Understanding*, *73*(1), 82–98.

Goldman, A. (1970). *A theory of human action*. Englewood Cliffs: Prentice Hall.

Gould, K., & Shah, M. (1989). The trajectory primal sketch: a multi-scale scheme for representing motion characteristics. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 79–85) San Diego, June 1989.

Gritai, A., & Shah, M. (2006). Tracking of human body joints using anthropometry. In *IEEE international conference on multimedia and expo*, Toronto, Canada, 2006.

Haritaoglu, I., Harwood, D., & Davis, L. (2000). W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 809.

Hartley, R. I., & Zisserman, A. (2000). *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press.

Herman, M. (1979). *Understanding body postures of human stick figures*. PhD Thesis, University of Maryland.

Hogg, D. C. (1984) *Interpreting images of a known moving object*. PhD thesis, University of Sussex.

Horn, B., & Weldon, E. (1988). Direct methods for recovering motion. *International Journal of Computer Vision*, *2*(1), 51.

Hu, W., Wang, L., & Tan, T. (2003). Recent development in human motion analysis. *Pattern Recognition*, *36*(3), 585.

Johansson, G. (1993). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, *14*(2), 201–211.

Johnson, N., Galata, A., & Hogg, D. (2001). Learning variable length Markov models of behaviour. *Computer Vision and Image Understanding Journal*, *81*, 398–413.

Ju, S., Black, M., & Yacoob, Y. (1996). Cardboard people: A parameterized model of articulated image motion. In *Proc. IEEE int. conf. on automatic face and gesture recognition* (pp. 38–44) 1996.

Kroemer, K., Easterby, R., & Chaffin, D. (1982). *Anthropometry and biomechanics—theory and application*. New York: Plenum Press.

Laptev, I. et al. (2005). Periodic motion detection and segmentation via approximate sequence alignment. In *Proceedings of the IEEE international conference on computer vision*, 2005.

Li, H., & Greenspan, M. (2005). Multi-scale gesture recognition from time-varying contours. In *Proceedings of the IEEE international conference on computer vision* (pp. 236–243) 2005.

Liao, W., Aggarwal, J., Cai, Q., & Sabata, B. (1994). Articulated and elastic non-rigid motion: a review. In *Workshop on motion of non-rigid and articulated objects*, 1994.

Moeslund, T., & Granumm, E. (2001). A survey of computer vision based human motion capture. *Computer Vision and Image Understanding*, *81*(3), 231.

Nishikawa, A., Ohnishi, A., & Miyazaki, F. (1998). Description and recognition of human gestures based on the transition of curvature from motion images. In *Proc. IEEE int. conf. on automatic face and gesture recognition* (pp. 552–557) 1998.

Niyogi, S., & Adelson, E. H. (1994). Analyzing and recognizing walking figures in xyt. In *IEEE international conference on computer vision and pattern recognition* (pp. 469–474) 1994.

O'Rourke, J., & Badler, N. (1980). Model-based image analysis of human motion using constraint propagation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *2*(6), 522.

Oliver, N., Rosario, B., & Pentland, A. (1999). A Bayesian computer vision system for modeling human interactions. In *Proceedings of ICVS99*, Gran Canaria, Spain, January 1999.

Oliver, N., Horvitz, E., & Garg, A. (2002). Layered representations for human activity recognition. In *Fourth IEEE int. conf. on multimodal interfaces* (pp. 3–8) 2002.

Parameswaran, V., & Chellappa, R. (2002). Quasi-invariants for human action representation and recognition. In *International conference on pattern recognition*, 2002.

Parameswaran, V., & Chellappa, R. (2003). View invariants for human action recognition. In *Proceedings of the IEEE international conference on computer vision and pattern recognition*, 2003.

Parameswaran, V., & Chellappa, R. (2009). Using 2d projective invariance for human action recognition. *International Journal of Computer Vision*.

Polana, R., & Nelson, R. C. (1994). Detecting activities. *Journal of Visual Communication and Image Representation*, *5*, 172–180.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, *9*(2), 129.

Rao, C., & Shah, M. (2001). View invariance in action recognition. In *Proceedings of the IEEE international conference on computer vision and pattern recognition*, Kauai, Hawaii, Dec. 2001.

Rao, C., Gritai, A., Shah, M., & Syeda-Mahmood, T. (2003). View-invariant alignment and matching of video sequences. In *Proceedings of the IEEE international conference on computer vision* (pp. 939–945) 2003.

Rashid, R. (1980). Towards a system for the interpretation of moving light display. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *2*(6), 574.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Spech, and Signal Processing*, *26*(1), 43.

Seitz, S. M., & Dyer, C. R. (1997). View-invariant analysis of cyclic motion. *International Journal of Computer Vision*, *25*, 1–25.

Shah, M., Gritai, A., & Sheikh, Y. (2004). On the use of anthropometry in the invariant analysis of human actions. In *International conference on pattern recognition*, 2004.

Shechtman, E., & Irani, M. (2005). Space-time behavior based correlation. In *IEEE conference on computer vision and pattern recognition*, June 2005.

Sheikh, Y., Gritai, A., & Shah, M. (2007). On the spacetime geometry of Galilean cameras. In *Proceedings of the IEEE international conference on computer vision and pattern recognition*, 2007.

Singer, Y., Fine, S., & Tishby, N. (1998). The hierarchical hidden Markov model: analysis and applications. *Machine Learning*, *32*(1), 41–62.

Starner, T., & Pentland, A. (1996). Motion-based recognition. In *Computational imaging and vision series. Real-time American sign language recognition from video using hidden Markov models*. Dordrecht: Kluwer Academic.

Sukthankar, R., Ke, Y., & Hebert, M. (2005). Efficient visual event detection using volumetric features. In *Proceedings of the IEEE international conference on computer vision*, Oct. 2005.

Venkatesh, S., Nguyen, N., Phung, D., & Bui, H. H. (2005). Learning and detecting activities from movement trajectories using the hierarchical hidden Markov models. In *Proceedings of the IEEE international conference on computer vision and pattern recognition*, San Diego, CA, 2005.

Verfaillie, K. (1992). Variant points of view on viewpoint invariance. *Canadian Journal of Psychology*, *46*, 215.

Von Mises, L. (1966). *Human action: a treatise on economics*. Chicago: Henry Regnery.

Yamato, J., Ohya, J., & Ishii, L. (1995). Recognizing human action in time-sequential images using hidden Markov model. In *Proc. of the IEEE conference on computer vision and pattern recognition* (pp. 624–630) 1995.

Yang, M., & Ahuja, N. (1998). Extracting gestural motion trajectories. In *Proc. IEEE int. conf. on automatic face and gesture recognition* (pp. 10–15) 1998.

Yang, J., Xu, Y., & Chen, C. S. (1997). Human action learning via hidden Markov model. *IEEE Transactions on System, Man, and Cybernetics*, *27*(1), 34–44.

Yilmaz, A., & Shah, M. (2005). Actions as objects: a novel action representation. In *IEEE Proceedings on the international conference on computer vision and pattern recognition*, 2005.

Zatsiorsky, V. (2002). *Kinematics of human motion*. Champaign: Human Kinetics.

Zelnik-Manor, L., & Irani, M. (2001). Event-based analysis of video. In *IEEE conference on computer vision and pattern recognition*, Dec. 2001.