2009

Geometric Invariance In The Analysis Of Human Motion In Video Data

Yuping Shen
University of Central Florida

GEOMETRIC INVARIANCE IN THE ANALYSIS OF HUMAN MOTION IN VIDEO DATA

by

YUPING SHEN
M.S. University of Central Florida, 2007
B.Eng. University of Science and Technology of China, 2004

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the School of Electrical Engineering and Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2009

Major Professor:
Hassan Foroosh

ABSTRACT

Human motion analysis is one of the major problems in computer vision research. It deals with the study of the motion of human body in video data from different aspects, ranging from the tracking of body parts and reconstruction of 3D human body configuration, to higher level of interpretation of human action and activities in image sequences. When human motion is observed through video camera, it is perspectively distorted and may appear totally different from different viewpoints. Therefore it is highly challenging to establish correct relationships between human motions across video sequences with different camera settings. In this work, we investigate the geometric invariance in the motion of human body, which is critical to accurately understand human motion in video data regardless of variations in camera parameters and viewpoints.

In human action analysis, the representation of human action is a very important issue, and it usually determines the nature of the solutions, including their limits in resolving the problem. Unlike existing research that study human motion as a whole 2D/3D object or a sequence of postures, we study human motion as a sequence of body pose transitions. We also decompose a human body pose further into a number of body point triplets, and break down a pose transition into the transition of a set of body point triplets. In this way the study of complex non-rigid motion of human body is reduced to that of the motion of rigid body point triplets, i.e. a collection of planes in

motion. As a result, projective geometry and linear algebra can be applied to explore the geometric invariance in human motion. Based on this formulation, we have discovered the fundamental ratio invariant and the eigenvalue equality invariant in human motion. We also propose solutions based on these geometric invariants to the problems of view-invariant recognition of human postures and actions, as well as analysis of human motion styles. These invariants and their applicability have been validated by experimental results supporting that their effectiveness in understanding human motion with various camera parameters and viewpoints.

This dissertation is dedicated to my parents and my wife who have been supporting me all the way.

ACKNOWLEDGMENTS

I would like to thank my advisor *Prof. Hassan Foroosh* for his valuable guidance, encouragement and friendship. Starting from my first steps into the Ph.D. program, his generous and sincere support helped me evolve from a student to a scientist.

I would like to thank *Prof. Charles Hughes*, *Prof. Michael Moshell*, and *Prof. Mashall Tappen* for serving as my committee members and for their valuable comments.

I would like to thank Yunjun Zhang, Fei Lu, Nazim Ashraf and Xiaochun Cao for their valuable help and friendship during my PhD study. I would also like to thank the previous and current members of the CIL lab, namely, Mais Alnasser, Adeel Buttha, Murat Balci, Imran Junejo, Brendan Moore, Remo Pillat, Alex Cook and Arun Kulshreshth for their help and friendship.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Human Motion Analysis

Human motion analysis is one of the major problems in computer vision research, and has received increasing attention from researchers. It has arisen from the need to detect and interpret human motion and activities in applications over a wide spectrum of topics, such as surveillance, kinesiology, video communications, entertainment industry, human computer interaction, etc. For example, in surveillance systems, human activities are monitored by computers in security-sensitive areas such as airports, borders, banks and building lobbies, etc. In kinesiology, the movement of body parts or joints are tracked and reconstructed in 3D space for the analysis of athletic performance and medical diagnostics. The capability to recognize different human actions is essential to the human computer interaction systems. Other applications include video conferences, digital libraries, etc.

The study of human motion is also a subject of interest in various disciplines such as philosophy [Gol70], psychology [Pri97], kinesiology [HS96], cognitive neurosciences [GFF96, BD01a], etc. In psychology, Johansson [JOH73] demonstrated that humans can identify motion from a small set of moving dots, from his classic experiments in which light displays are attached to various body parts. In kinesiology the goal is to study human movements, performance and function, and to

preserve and enhance human movements with applications in sports, medical care, recreation, and exercise.

In computer vision, the goal of human motion analysis is to study the motion of human body from image sequences in different aspects, ranging from the tracking of body parts and reconstruction of 3D human body configuration, to higher level of interpretation of human action and activities from image sequences. Different levels of tasks are usually involved in the analysis of human motion from video data, including *human detection, body parts/joints tracking, human posture recognition, human body configuration recovery, human action recognition, human activities recognition, and human action style analysis.* In this thesis, we focus on three major areas: (1) human posture recognition, (2) human action recognition and (3) human action style analysis. In the following sections, we give a brief introduction to these problems, and more detailed discussion can be found in chapters 3 and 4.

1.1.1 Recognition of Human Postures

The goal of human postures recognition is to recognize a pre-defined human body pose in a video sequence or an image. It is a fundamental problem in many applications such as visual surveillance, video summarization and video indexing, etc. The representation of human pose is critical to the task of pose recognition, and determines the nature of its solutions. Model based approaches use a 3D point-based model of human body pose, and recognize the pose in 2D images by fitting the 3D

model to the 2D appearance of body pose on the images. Another representation is based on 2D silhouettes or body joints of the body pose. Since depth information is lost in this representation, multiple views are usually required for view-invariant pose recognition, although it is possible to achieve view invariance from one single view, as discussed here, as well as in [PC03, YS06, SS05]. When the posture is studied in the context of video sequence, temporal information is also considered as a feature to represent the body pose, yielding a spatio-temporal representation of human body pose. One example of such representation is the template based spatio-temporal representation proposed in this thesis, which is based on transition of poses in a video sequence.

The problem of pose recognition has proven to be a significant challenge, partially due to the high degrees of freedom of human body structure. As demonstrated in [Zat02], the human body has no less than 244 degrees of freedom, which leads to high dimensional search space in the model based approaches. Another source of challenge is the camera parameters and viewpoint variations. After perspective projection to camera images, the same body pose may appear totally different when observed from different viewpoints, which makes it rather difficult to build a relationship between two different views.

1.1.2 Human Action Recognition

Action recognition could be regarded as an extension of the pose recognition problem, to include the dynamics of human body. It is a challenging problem that combines the uncertainty associated

with computational vision and unpredictable human behavior. The high non-rigidity of human body pose makes the task of modeling the dynamics of body pose extremely difficult. In addition, based on research in anthropology [Far99], the body dynamics have been found to be affected by many factors, including age, ethnicity, class, gender, circumstance, etc. Like pose recognition, viewpoint variations is also critical to action recognition. It is unreasonable and very limited to assume that the viewpoints remain constant across different observations of an action. Hence, it is important that an action recognition system be independent of camera internal calibration parameters and viewpoint changes. The problem of view invariance has attracted an increasing attention of researchers, and is a crucial problem in the state-of-the-art research of action recognition. Temporal variability of action sequences is another source of challenges in action recognition. Different individuals may execute the same action at different rates, except for some rare scenarios such as synchronized dancing. Furthermore, the camera frame-rates may also vary in different observations of the same action. These factors lead to a dynamic time-line mapping between different observations of an action, which requires action recognition algorithms to be tolerant of high temporal variability.

The representation of human action is also of great importance to action recognition. Existing representations include: a sequence of body poses, in the form of body joints or silhouettes; 3D space-time objects of the entire action sequence, such as space-time trajectories of body joints and space-time cylinder or volume of silhouettes; and 2D history template of silhouettes. In this thesis, we introduce a novel presentation of human action, i.e., a sequence of pose transitions, in the form of triplet motions. We break down a point-based model of a body pose into body point triplets,

and describe the transition of two body poses by the rigid motion of planes determined by these triplets.

1.1.3 Human Action Style Analysis

Unlike human action recognition that discriminates different types of human motion such as jumping, dancing and walking, the purpose of human action style analysis is to study the variations of individuals in performing the same action. This is motivated by the fact that human actions exhibits certain styles determined by many factors such as human identity, gender, age, environments, etc. For example, the walking pattern of a person is usually performed in a fairly repeatable and unique way or style, which makes it possible to recognize a person at a distance by their gait. People can also easily perceive whether a person is walking ”in a hurry” or just ”casually strolling along”. It is also possible to determine whether a golf player is professional or amateur by looking at the style of his/her swing. Furthermore, from the style differences of walking, people can recognize its internal physical origins such as age [Dav01], and gender [BCK78, CPK78, KC77, Tro02]. Environment variations could also be determined by studying the walking styles of people, such as ”walking on uneven ground”, ”carrying a heavy suitcase”, etc. The goal of human action style analysis is to represent and recognize such action styles, and extract underlying factors that determine their differences.

The study of action style is motivated by the need of various applications. Automatic human identification from gait is a desired capability of advanced surveillance systems, especially when other biometrics such as iris and finger prints are not perceivable at a distance. Rather than just reporting ”moving objects” or ”walking person”, it is also desirable to generate more qualitative motion descriptions of human behavior such as ”walking in a hurry” or ”walking person carrying heavy objects”, which may help the system to detect irregular behaviors of persons. Another application of human action style analysis is ergonomic evaluation, for instance, for detecting improper techniques of athletes to help reduce the occurrence of injury. It can be also applied in the re-use of motion capture data, by translating human action animations into new styles [BH00, DK02]. In this thesis, we study the triplet motion variations and pose variations in the same class of action, and extract stylistic features based on our geometric invariants, and apply them to resolve the gender recognition problem.

1.2 Geometric Invariance in Human Motion Analysis

In computer vision, data are acquired in forms of video sequences or images through cameras. In the process of projecting 3D scenes to 2D images, depth is lost and objects are perspectively distorted. Consequentially, the same objects may appear totally different when observed from different cameras, especially from different viewpoints, which makes it very challenging to find the geometric relationship in different camera images. In return, the lost depth information could be recovered from the analysis of geometric relationship between multiple views of the object, which

makes many tasks such as 3D reconstruction become possible. Hence, geometry is an important and ubiquitous tool in computer vision.

Geometric reasoning, which is to analyze the geometric relationships inherent in the perspective projection of objects from 3D space to camera image, plays an important role in the research in computer vision, and underpins the development of many working systems for tasks such as 3D graphics modelling, object recognition, scene analysis, video augmentation, etc.

As one of the fundamental tasks in geometric reasoning, the study of geometric invariance has arisen from the desire to identify geometric properties of objects that are invariant to the camera internal calibration parameters and the viewpoint from which the image is acquired. It is deeply rooted in the study of algebraic and projective invariants, which is a classical problem in mathematics dating back to the 18th century [Hil93]. There are two types of geometric invariance: multiple view invariants and single view invariants.

1.2.1 Multiple View Invariants

The analysis of multiple views is mostly based on the notion of epipolar geometry, which relates corresponding image points across views by a fundamental matrix. The projection of a 3D point in one view determines an epipolar line in any other view, on which the corresponding projected point lies. Such epipolar lines intersect at an epipole, which is the image of the camera center of the first view in the other cameras. These properties do not depend on camera internal parameters

or viewpoint, and always hold for any two views with different camera centers. Epipolar geometry is widely used in many applications, such as 3D reconstruction [CBP05, FCZ98], video editing [RPL05, BB87], action recognition [YS06, SS05], etc.

Another direction to analyze multiple views is based on the machinery of algebra. Carlsson [Car94] applies bracket algebra to the analysis of multiple views and shows that the resultant framework can be used to derive classical multiple view invariants. A similar work which uses algebraic elimination was proposed by Barrett [BGP93]. Weiss [Wei94] proposed a framework to discover new multiple view invariants, which is based on methods taken from the invariants of physical laws.

1.2.2 Single View Invariants

It has been proven in [MUT91] that, when a 3D set of points are in general position, i.e., completely unconstrained, invariants cannot be measured from a single view. However, single view invariants are available if the structure of the 3D points is constrained, e.g., surfaces of revolution [CBP05], objects with bilateral symmetry and polyhedra [ST93], straight homogeneous generalized cylinders [ZN94], etc.

A special scenario in single view analysis is the existence of repeated structures, which provides multiple view information from one single view. For instance, Zhang et al. [ZT98] use plane mirrors to capture multiple views of objects for 3D reconstruction. Another example is that two

identical objects in a single view are related by a translation. It is equivalent to a stereo pair of images taken from cameras related by a pure translation. In such cases, the multiple view invariants can be reinterpreted in the single view domain.

1.2.3 Geometric Invariance in Recognition of Objects and Human Motion

A critical problem of object recognition is to find ideal image descriptors that capture essential traits of an object and are insensitive to environmental changes such as illumination variations. The perspective projection and viewpoint variations cause great variability in object appearance in images, which requires the geometric invariance of image descriptors. Indeed, the nature of the problem of object recognition is the search for invariants [Wei93], among which the geometric and illumination invariants are recognized as most important.

The application of geometric invariants have been studied extensively in the literature of object recognition. Invariants under affine transformation have been studied in [ASB90, KH94, MW97], in which affine invariants were used to recognize planar objects in 3D space. Lei [Lei90] formulated perspective invariants based on cross ratios to recognize polygonal planar objects. Invariants for recognizing objects of more complex shape have also been studied. Invariant representations that use flat lists of invariant values computed on isolated portions or simple sequences of object boundary have been widely used to recognize objects of arbitrary shape. Subrahmonia et al. [SCK96] use planar shapes to approximate high order algebraic curves, and derive shape invari-

ants based on a subspace of the algebraic coefficients. Rothwell [FR94] uses invariant geometric relations to build a hierarchical description of objects. More detailed discussion of invariants in object recognition could be found in [CF01].

Another application of geometric invariants is in scene analysis, especially in human motion analysis, which is the subject of this thesis. Unlike objects of arbitrary shape, the structure of human body is more constrained, which, however, does not simplify the task of recognition, partly due to the high degree of freedom of the articulated human body, and the irregular dynamic of the structure. Since human motion can be regarded as 4D data with 3D Euclidean space and a temporal dimension, the temporal information is inherent in the study of geometric invariants in human motion analysis. For example, Rao et al. [RYS02] described the concept of a dynamic instant which is based on curvature of the spatio-temporal trajectory of human body points. Seitz and Dyer [SD97] used view-invariant measurement to find the repeating pose of walking people and the re-occurrence of position of turning points. Laptev [LBP05] proposed using spatio-temporal points from the video to compute the fundamental matrix/homography, which are in temporal matrix format, to detect the periodic motion once the transformation between video clips are obtained. Yilmaz [YS06] demonstrate that a temporal fundamental matrix relates two sequences of similar human action, and can be used to design a system for view-invariant action recognition. Parameswaran and Chellappa [PC03] 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. Here we exploit two types of view invariants in the transition of two poses in an action

sequence and use them for various applications including pose and action recognition, and human style analysis.

1.3 Organization of the Dissertation

The rest of the dissertation is organized as follows:

In chapter 2, we discuss a representation of human body pose, and present the technique to decompose a human posture into a set of body point triplets. We follow by proposing two types of geometric invariants in analyzing triplet motion based on fundamental ratios in section 2.3, and eigenvalues equality of homology in section 2.4. We describe the human motion by the motion of point triplets, and hence these two geometric invariants on moving triplets provide us a means of measuring the similarity between two pose transitions, which is discussed in section 2.5.

In chapter 3, we demonstrate the application of proposed geometric invariants to view-invariant recognition of human pose and action. We propose a template based technique for human pose recognition in section 3.1. Our templates for specific body poses are extracted from video data, and in addition to spatial information encode temporal information of body pose transitions. In section 3.2 we present the our solution to view-invariant human action recognition using our geometric invariants. We represent an action sequence by a sequence of pose transitions, unlike the existing methods that regard it as a whole or a sequence of poses. We then propose a dynamic programming based algorithm to solve the action recognition. Our methods are evaluated by experiments on

semi-synthetic and real video data in section 3.2.4, which show that our methods can recognize human poses and actions under substantial amount of noise, even when the viewpoints and camera parameters are unknown and totally different.

In chapter 4, we propose extensive work in the human motion analysis. We first discuss the contribution of different body point triplets in the task of pose and action recognition, and propose a technique to enhance the performance of our pose and action recognition algorithms. In chapter 5, we study the triplet motion variations and pose variations in the same class of action, and utilize them as representations of the stylistic information of human actions. We extract stylistic features from these representations based on principal component analysis (PCA) and Fisher Linear Discriminant Analysis (FLDA), and apply them in the gender recognition problem.

2 GEOMETRIC INVARIANCE IN MATCHING POSE TRANSITION

Depending on the nature and the scale of the problem, understanding of human motion is usually studied at four levels of granularity in the literature:

1. Pose : Pose, which is a snapshot of a person's posture, is the most atomic primitive of human motion. Although a single pose is stationary, together with contextual information, it can convey human motion or action. For instance, one can recognize a person playing golf from a single pose shown in Fig. 2.2. Given some prior knowledge, additional recovery of 3D structure of a person is also possible. This is essentially the underlying motivation for research on human pose estimation using a single image and contextual information [MM02, RS00a]. On the other hand, since an action may be regarded as a sequence of poses, determining the similarity between poses is the key component of many action recognition techniques.

2. Movement : Movement is a temporal fragment of human action, which can be regarded as the stacking of several poses in time. Movements are short segments of an action, and require

no contextual knowledge to be recognized. Bending knee, raising left hand, lowering body, etc., are some examples of movements.

3. Action : An action, e.g., walking, running, sitting down, is composed of an ordered set of movements. The temporal order of movements is predefined for each action, while their temporal rates may vary in the same action performed by different individuals. For example, the action "breast-stroke swim" is composed of the following movements in order: pull legs forward, extend legs, sweep arms, lift head and shoulders up.

4. Activity : An activity is composed of actions of individuals, as well as the interactions with the environment and between individuals. For instance, the activity "playing a football game" consists of various actions performed by a number of players, including pushing, running, kicking, etc.

Movement described above can be regarded as a shorter version of action, and the analysis of movement is actually as hard as analyzing actions. However, there is another level between movement and stationary pose, that is *pose transition*. The transition between two poses describes the motion of body parts. It holds the temporal information of human motion, while keeping the task at the atomic level. The study of pose transition can also be easily extended to higher level such as action recognition, due to the fact that an action can be regarded as a sequence of pose transition. With these properties pose transition plays an important role in the analysis of human motion. As shown in the rest of this thesis, the use of pose transition is the basis of our human motion analysis system.

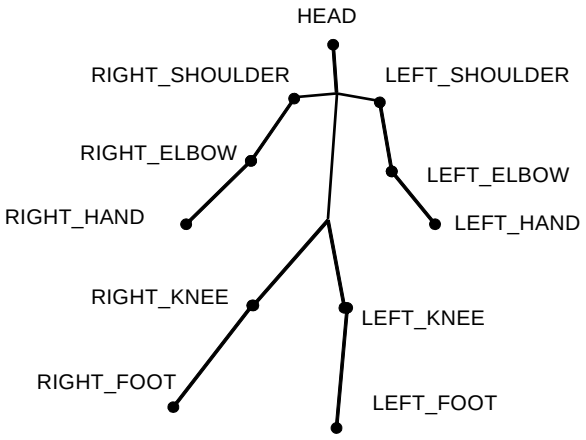


Figure 2.1: Human body model used in this thesis with 11 body points: head, shoulders, elbows, hands, knees and feet.

In this chapter, we focus on pose transition and aim at finding geometric invariants for the fundamental problem of pose transition matching. The resulting theory could be applied to extended research in pose recognition (see section 3.1, action recognition (see section 3.2) and human motion style analysis (see 4). In the following sections, we first discuss some fundamental issues in analyzing human motion, and then present two types of geometric invariants in pose transition matching.

2.1 Representation of Human Body Pose

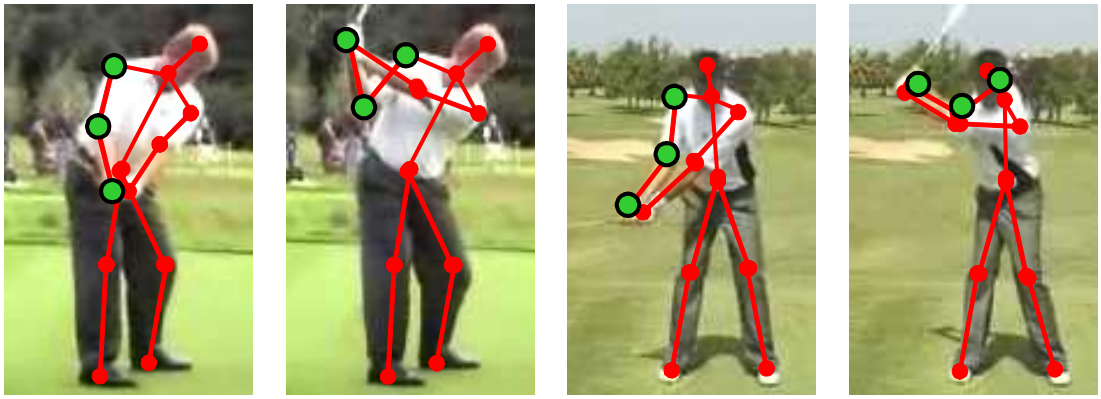
Set of body points is a widely used representation in action recognition, partly due to the fact that human body can be regarded as an articulate object. Using this presentation, an action is represented as a sequence of point sets, or a set of point trajectories in time. As reported by

Johansson [JOH73] in his classic experiments, human can identify motion when presented with only a small set of moving dots attached to various body parts, which has also been supported by recent research on human motion analysis, such as [GSS04, JOH73, PC03, YS06]. Other representations of pose include subject silhouette [BGS05, BD01b, SVS01], optical flow [EBM03, Wan06, ZXG06] and local space time features [LBP05, SLC04].

In this thesis, we use a body model which is represented by the image of 11 joints and end points, including head, shoulders, elbows, hands, knees and feet (see Fig. 2.1). These points, which are the only inputs to our algorithm, can be obtained by using articulated object tracking techniques such as [PRC99, RK95]. Further discussions on articulated object tracking can be found in [AC99, Gav99, MG01], and is beyond the scope of this thesis. We shall, henceforth, assume that tracking has already been performed on the data, and that we are given a set of labelled points for each image.

2.2 Human Body Pose Decomposition

Despite the fact that the articulated human body model shown in figure 2.1 has large degrees of freedom, the body points are not independent of each other. They are constrained by the structure of human body, and interact with each other during the execution of a human motion. Taking into consideration of such properties, we decompose this body model into triplets of points, rather than studying the body point set as a whole. Since any three non-collinear points in the 3D space



I_1 I_2 J_i J_j

Figure 2.2: An example of similar pose transitions. The transition from I_1 to I_2 is similar to that from J_i to J_j. The pose transition can be broken down into a set of moving triplets of points, e.g., the highlighted triplet of points in the images.

.

define a planar surface, this effectively breaks down the articulated human body into a collection of planar surfaces defined by every non-degenerate triplet in the image plane. When a body pose is studied in the context of a video sequence, temporal information also provides useful information for identifying the pose. Our template encodes temporal information in the form of pose transition as described earlier. We can thus match a pose pair $\langle I_i, I_j \rangle$ and a template $\langle T_1^k, T_2^k \rangle$ by comparing their pose transitions[1]. Using our point-triplet representation has the following advantages:

- The similarity of pose transitions for articulated bodies can be measured by matching the rigid motions of scene planes defined by all triplets of body points.

[1]For brevity of notation, we will hereafter denote a pair $\langle I_i, I_j \rangle$ as $I_{i,j}$.

- The representation leads to a highly over-determined formulation of the problem, allowing us to achieve robustness to noise and self-occlusions: Given n body point correspondences, we obtain $\binom{n}{3}$ criteria to measure the similarity. Even if there exist occluded body points, they would be inconsequential since the vast majority of point triplets would be typically available to fulfill the task.

- Anthropometric restrictions can be relaxed, since only the transitions of planes in the 3D space matter, and not the points defining these planes or the ratios of the distances between these points.

2.3 Fundamental Ratios Invariant

Now we consider the problem of matching two pose transitions. Using the triplet representation of human body pose, the problem is reduced to matching motions of corresponding body point triplets. In this section, we present a geometric invariant of triplet motion based on fundamental ratios, which can be used to measure the similarity of triplet motion and, furthermore, the pose transition.

2.3.1 Fundamental Ratios

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

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

$$\mathbf{F}^{2\times 2} \sim \begin{bmatrix} \epsilon_{1st}\mathbf{t}^s\mathbf{r}_1^t & \epsilon_{1st}\mathbf{t}^s\mathbf{r}_2^t \\ \epsilon_{2st}\mathbf{t}^s\mathbf{r}_1^t & \epsilon_{2st}\mathbf{t}^s\mathbf{r}_2^t \end{bmatrix}, \tag{2.1}$$

where $\mathbf{r}_k$ *is the* k*-th column of* $\mathbf{R}$*, the superscript, e.g.* i*, refers to* i^{th} *element of a vector, and* ϵ_{rst} *for* $r, s, t = 1, \ldots, 3$ *is a permutation tensor*[2].

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

The upper 2×2 sub-matrices $\mathbf{F}^{2\times 2}$ for two moving cameras could be used to measure the similarity of camera motions. That is, if two cameras perform the same motion (same relative translation and

[2] *The use of tensor notation is explained in details in [HZ04], p563.*

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

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

$$\mathbf{F}_1^{2\times2} \sim \mathbf{F}_2^{2\times2} \tag{2.2}$$

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

Here similar motion implies that plane normals undergo same motion up to a similarity transformation. The projective nature of the view-invariant equation in (2.2) implies that the elements in the sub-matrices on the both sides of (2.2) are equal up to an arbitrary non-zero scale factor, and hence only the ratios among them matter. We call these ratios the *fundamental ratios*, and as propositions 1 and 2 imply, these fundamental ratios are invariant to camera intrinsic parameters and viewpoint. To eliminate the scale factor, we can normalize both sides using $\hat{\mathbf{F}}_i = |\mathbf{F}_i^{2\times2}|/\|\mathbf{F}_i^{2\times2}\|_F, i = 1, 2$, where $|\cdot|$ refers to absolute value operator and $\|\cdot\|_F$ stands for the Frobenius norm. We then have

$$\hat{\mathbf{F}}_1 = \hat{\mathbf{F}}_2 \tag{2.3}$$

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

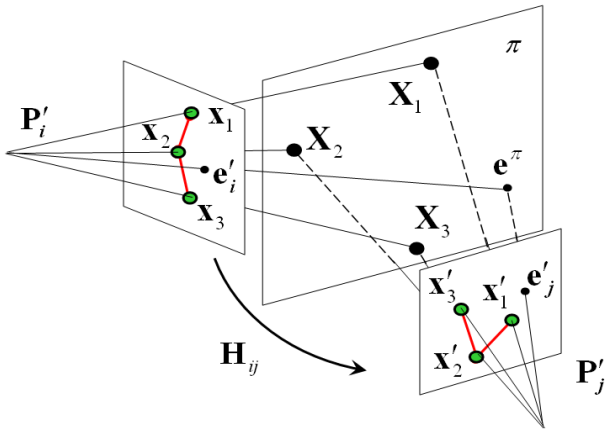


Figure 2.3: Fundamental matrix induced by a moving plane is dual to a stationary plane with moving camera.

the residual error:

$$E_f(\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2) = \|\hat{\mathbf{F}}_1 - \hat{\mathbf{F}}_2\|_F \tag{2.4}$$

2.3.2 Constraints on Inter-pose Fundamental Matrices Induced by Moving Triplets

Here we describe how to apply the theoretic results derived in section 2.3.1 to our problem. We are given an observed pose transition $I_i \rightarrow I_j$ from sequence $\{I_t\}$, and a second one $J_m^k \rightarrow J_n^k$ from sequence $\{J_t^k\}$. When $I_i \rightarrow I_j$ corresponds to $J_m^k \rightarrow J_n^k$, one can regard them as observations of the same 3D pose transition from two different cameras $\mathbf{P}_1$ and $\mathbf{P}_2$, respectively. There are two instances of epipolar geometry associated with this scenario:

1. The mapping between the image pair $\langle I_i, I_j \rangle$ and the image pair $\langle J_m^k, J_n^k \rangle$ is determined by the fundamental matrix $\mathbf{F}$ [HZ04] related to $\mathbf{P}_1$ and $\mathbf{P}_2$. The projection of the camera center

of $\mathbf{P}_2$ in I_i or I_j is given by the epipole $\mathbf{e}_1$, which is found as the right null vector of $\mathbf{F}$. Similarly the image of the camera center of $\mathbf{P}_1$ in J_m^k or J_n^k is the epipole $\mathbf{e}_2$ given by the right null vector of $\mathbf{F}^T$.

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

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

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

Δ_i and Δ_j can be regarded as projections of a stationary 3D point triplet $\langle \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3 \rangle$ on two virtual cameras $\mathbf{P}'_i$ and $\mathbf{P}'_j$, as shown in Fig. 2.3. $\langle \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3 \rangle$ defines a world plane π, which induces a homography $\mathbf{H}_{ij}$ between $\mathbf{P}'_i$ and $\mathbf{P}'_j$. It is known that a homography may be computed from four corresponding image points. In this case, the four points can be the image points $\mathbf{x}_1, ..., \mathbf{x}_3$ and $\mathbf{x}'_1, ..., \mathbf{x}'_3$ together with the epipoles in $\mathbf{P}'_i$ and $\mathbf{P}'_j$. Let $\mathbf{e}'_i$ and $\mathbf{e}'_j$ be these epipoles. If $\mathbf{e}'_i$ and $\mathbf{e}'_j$ are known, then $\mathbf{H}_{ij}$ can be computed, and hence $\mathcal{F}_1$ induced by Δ_i and Δ_j can be determined using

$$\mathcal{F}_1 = [\mathbf{e}'_j]_\times \mathbf{H}_{ij}, \text{ or } \mathcal{F}_1 = \mathbf{H}_{ij}^{-T} [\mathbf{e}'_i]_\times. \tag{2.5}$$

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

$$\mathcal{F}_2 = [\mathbf{e}'_n]_\times \mathbf{H}_{mn}, \text{ or } \mathcal{F}_2 = \mathbf{H}_{mn}^{-T}[\mathbf{e}'_m]_\times, \tag{2.6}$$

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

The difficulty with (2.5) and (2.6) is that the epipoles $\mathbf{e}'_i$, $\mathbf{e}'_j$, $\mathbf{e}'_m$ and $\mathbf{e}'_n$ are unknown, and cannot be computed directly from the triplet correspondences. Fortunately, however, the epipoles can be closely approximated as described below.

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

$$\mathbf{e}'_i = \mathbf{e}'_j = \mathbf{e}_1, \quad \mathbf{e}'_m = \mathbf{e}'_n = \mathbf{e}_2, \tag{2.7}$$

we have:

$$E_f(\hat{\mathcal{F}}_1, \hat{\mathcal{F}}_2) = 0. \tag{2.8}$$

Under more general motion, the equalities in (2.7) become only approximate. However, we shall see in section 2.6.1 that this approximation is inconsequential in action recognition for a wide range of practical rotation angles. Using equation (2.4) and the fundamental matrices $\mathcal{F}_1$ and $\mathcal{F}_2$ computed for every non-degenerate triplet, we can define a similarity measure for matching pose transitions $I_i \rightarrow I_j$ and $J_m^k \rightarrow J_n^k$:

$$\mathcal{E}_f(I_i \rightarrow I_j, J_m^k \rightarrow J_n^k) = \underset{\text{all } \Delta_i}{\text{Median}}(E_f(\hat{\mathcal{F}}_1, \hat{\mathcal{F}}_2)). \tag{2.9}$$

2.3.3 Degenerate Triplets

A homography cannot be computed from four correspondences if three points are collinear. Even when three image points are close to collinear the problem becomes ill-conditioned. We call such triplets as degenerate, and simply ignore them in matching pose transitions. This does not produce any difficulty in practice, since with 11 body point representation we use (see Fig. 2.1), we obtain 165 possible triplets, the vast majority of which are in practice non-degenerate. A special case is when the epipole is close to or at infinity, for which all triplets would degenerate.

We solve this problem by transforming the image points in projective space, which is motivated by Zhang et al. [ZL01]. The idea is to find a pair of projective transformations $\mathbf{Q}$ and $\mathbf{Q}'$, such that after transformation the epipoles and transformed image points are not at infinity. Suppose corresponding image points $\{\mathbf{m}_i\}$ and $\{\mathbf{m}'_i\}$ are related by the fundamental matrix $\mathbf{F}$, with epipoles $\mathbf{e}$ and $\mathbf{e}'$. Then $\mathbf{Q}$ and $\mathbf{Q}'$ are determined as follows:

1. Normalize $\{\mathbf{m}_i\}$ and $\{\mathbf{m}'_i\}$ so that the centroid of each point set is at the origin and the RMS distance of the points from the origin is $\sqrt{2}$. Then translate both sets of points so that their Euclidean image coordinates lie in the region $\{x \geq 1, y \geq 1\}$. The combined transformations are $\mathbf{T}$ and $\mathbf{T}'$, and the fundamental matrix and the epipoles are:

$$\tilde{\mathbf{F}} = \mathbf{T}'^{-T}\mathbf{F}\mathbf{T}^{-1},\ \tilde{\mathbf{F}}\tilde{\mathbf{e}} = 0,\ \text{and}\ \tilde{\mathbf{F}}^T\tilde{\mathbf{e}}' = 0. \tag{2.10}$$

2. Initialize $\mathbf{Q}$ and $\mathbf{Q}'$ to be identity matrices.

3. Let $\tilde{\mathbf{e}}[i]$ be the i^{th} $(i = 1, 2, 3)$ element of $\tilde{\mathbf{e}}$. If $|\tilde{\mathbf{e}}[1]| > |\tilde{\mathbf{e}}[3]|$, permute elements 1 and 3 of $\tilde{\mathbf{e}}$ and permute rows 1 and 3 of $\mathbf{Q}$. Then, if $|\tilde{\mathbf{e}}[2]| > |\tilde{\mathbf{e}}[3]|$, permute elements 2 and 3 of $\tilde{\mathbf{e}}$ and permute rows 2 and 3 of $\mathbf{Q}$.

4. If $|\tilde{\mathbf{e}}'[1]| > |\tilde{\mathbf{e}}'[3]|$, permute elements 1 and 3 of $\tilde{\mathbf{e}}'$ and permute rows 1 and 3 of $\mathbf{Q}'$. Then, if $|\tilde{\mathbf{e}}'[2]| > |\tilde{\mathbf{e}}'[3]|$, permute elements 2 and 3 of $\tilde{\mathbf{e}}'$ and permute rows 2 and 3 of $\mathbf{Q}'$.

5. $\mathbf{Q} = \mathbf{QT}$ and $\mathbf{Q}' = \mathbf{Q}'\mathbf{T}'$.

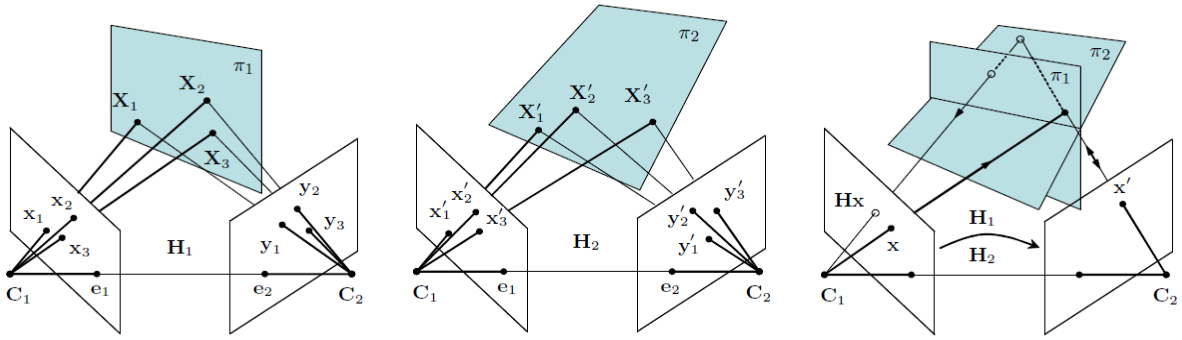
Steps 2-4 guaranty that after transformation the epipoles fall in the region $x, y \in [-1, 1]$, while step 1 ensures finite image points during permutations in steps 2-4. The fundamental matrix and the epipoles of the transformed points are

$$\tilde{\mathbf{F}} = \mathbf{Q}'^{-T}\mathbf{F}\mathbf{Q}^{-1},\ \tilde{\mathbf{F}}\tilde{\mathbf{e}} = 0,\ \text{and}\ \tilde{\mathbf{F}}^T\tilde{\mathbf{e}}' = 0. \tag{2.11}$$

Note that these transformations do not affect the projective equality in Proposition 2.

2.4 Eigenvalues Equality Invariant

Though fundamental ratios invariant described in section 2.3 can cover a wide range of scenarios in practice, it is still limited when hard viewpoints are encountered, and is not a "true" view invariant. In this section, we propose another view invariant that based on homology eigenvalues equality, and unlike the fundamental ratios invariant, it is valid from all viewpoints.



(a) (b) (c)

Figure 2.4: Homographies induced by a moving triplet of points. Suppose that the motion of a triplet of 3D points $\{\mathbf{X}_i\} \rightarrow \{\mathbf{X}'_i\}$ is observed by two stationary cameras, $\mathbf{C}_1$ and $\mathbf{C}_2$, as $\{\mathbf{x}_i\} \rightarrow \{\mathbf{x}'_i\}$ and $\{\mathbf{y}_i\} \rightarrow \{\mathbf{y}'_i\}$. Together with the epipoles $\mathbf{e}_1 \leftrightarrow \mathbf{e}_2$, the point correspondences $\mathbf{x}_i \leftrightarrow \mathbf{y}_i$ and $\mathbf{x}'_i \leftrightarrow \mathbf{y}'_i$ induce two homographies $\mathbf{H}_1$ and $\mathbf{H}_2$ from the left view to the right view. A homography that maps the left view to itself is then defined as $\mathbf{H} = \mathbf{H}_2^{-1}\mathbf{H}_1$. For similar motions of triplets, this homography is shown to be a homology and hence with two identical eigenvalues, providing thus a constraint for identifying similar pose transitions (see text for more details).

2.4.1 Homographies Induced by A Triplet of Body Points

Consider the case that $\langle I_1, I_2 \rangle$ corresponds to $\langle J_i, J_j \rangle$, and the transformation from I_1 to I_2 corresponds to that from J_i to J_j. $I_{1,2}$ and $J_{i,j}$ can then be regarded as the images of same moving subject viewed by two different cameras. Suppose that $I_{1,2}$ are observed by camera P_1 and $J_{i,j}$ by camera P_2. $\mathbf{P}_1$ and $\mathbf{P}_2$ may have different intrinsic and extrinsic parameters. As described earlier, these point correspondences induce an epipolar geometry via the fundamental matrix $\mathbf{F}$. The projection of the camera center of $\mathbf{P}_2$ in I_1 and I_2 is given by the epipole $\mathbf{e}_1$, which is found as the

right null vector of $\mathbf{F}$. Similarly the image of the camera center of $\mathbf{P}_1$ in J_i and J_j is the epipole $\mathbf{e}_2$ given by the right null vector of $\mathbf{F}^T$. Note that this fundamental matrix does not correlate the entire scene, but only the body points of the subjects.

Let us now consider an arbitrary triplet of 3D body points (see highlighted points in Fig. 2.2 for example), $\mathbf{\Delta} = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$, which corresponds to $\Delta_1 = \langle \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \rangle$ in I_1 and $\Delta_i = \langle \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3 \rangle$ in J_i. After the pose transformation, $\mathbf{\Delta}$ transforms to $\mathbf{\Delta}' = \{\mathbf{X}'_1, \mathbf{X}'_2, \mathbf{X}'_3\}$, which corresponds to $\Delta_2 = \langle \mathbf{x}'_1, \mathbf{x}'_2, \mathbf{x}'_3 \rangle$, in I_2 and $\Delta_j = \langle \mathbf{y}'_1, \mathbf{y}'_2, \mathbf{y}'_3 \rangle$ in J_j, as illustrated in Fig. 2.4.

$\mathbf{\Delta}$ and $\mathbf{\Delta}'$ determine two scene planes π_1 and π_2 in the 3D space, which induce two homographies $\mathbf{H}_1$ and $\mathbf{H}_2$ between $\mathbf{P}_1$ and $\mathbf{P}_2$. These plane-induced homographies can be computed given four point correspondences, i.e. the image point correspondences $\mathbf{x}_i \leftrightarrow \mathbf{y}_i$ and the epipoles $\mathbf{e}_1 \leftrightarrow \mathbf{e}_2$ by solving the system of linear equations:

$$\mathbf{y}_i \sim \mathbf{H}_1 \mathbf{x}_i \tag{2.12}$$

$$\mathbf{e}_2 \sim \mathbf{H}_1 \mathbf{e}_1 \tag{2.13}$$

where, as is customary, $\sim$ indicates projective equality up to an unknown scale. A similar set of equations provide $\mathbf{H}_2$:

$$\mathbf{y}'_i \sim \mathbf{H}_1 \mathbf{x}'_i \tag{2.14}$$

$$\mathbf{e}_2 \sim \mathbf{H}_1 \mathbf{e}_1 \tag{2.15}$$

2.4.2 Constraints on Homographies Induced by Moving Triplets

During a pose transition, the motion of a triplet $\boldsymbol{\Delta} \rightarrow \boldsymbol{\Delta}'$ specifies a rigid motion of a scene plane $\pi_1 \rightarrow \pi_2$, which induces two homographies $\mathbf{H}_1$ and $\mathbf{H}_2$. These homographies define a mapping from I_1 (or I_2) to itself given by

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

As shown in Fig. 2.4, $\mathbf{H}$ first maps a point $\mathbf{x}$ on I_1 (or I_2) to $\mathbf{y}$ on J_i (or J_j) through π_1, and then transforms it back to I_1 (or I_2) as $\mathbf{Hx}$ through π_2. It can be readily verified either algebraically or from Fig. 2.4 that points on the intersection of π_1 and π_2 are fixed during the mapping. Another fixed point under this mapping is the epipole $\mathbf{e}_1$. Thus the homography $\mathbf{H}$ has a line of fixed points (the intersection line of π_1 and π_2) and a fixed point not on the line (the epipole $\mathbf{e}_1$), which means that

Proposition 4 *If a triplet of 3D points observed by two cameras undergo the same motion, then the homography* $\mathbf{H}$ *reduces to a planar homology, and hence two of its eigenvalues must be equal.*

The equality of the two eigenvalues of $\mathbf{H}$ defines a consistency constraint on $\mathbf{H}_1$ and $\mathbf{H}_2$, imposing the assumption that the two cameras are observing the same scene plane motions. In the special case when the triplet is stationary, i.e., $I_1 = I_2$ and $J_i = J_j$, this equality constraint is still satisfied since $\mathbf{H}$ reduces to an identity matrix, with all its eigenvalues equal to 1. In practice, due to noise and subject-dependent differences, this constraint of equality of two eigenvalues for the

same pose transition can be expressed by defining the following error function on $\mathbf{H}$:

$$E_h(\mathbf{H}) = \frac{|a-b|}{|a+b|}, \tag{2.16}$$

where a, and b are the two closest eigenvalues of $\mathbf{H}$. $E_h(\mathbf{H})$ can be used to measure the similarity of motion of a triplet between two sequences, and the combination of $E_f(\mathbf{H})$ for all triplets of non-collinear points provides a measure of similarity between *pose transitions* $I_1 \rightarrow I_2$ and $J_i \rightarrow J_j$:

$$\mathcal{E}_h(I_1 \rightarrow I_2, J_i \rightarrow J_j) = \underset{\text{all } \Delta_i}{\text{Median}}(E_h(\mathbf{H})). \tag{2.17}$$

$\mathcal{E}_h(I_1 \rightarrow I_2, J_i \rightarrow J_j)$ is minimal for similar pose transitions, and is invariant to camera calibration matrix and viewpoint variations.

2.4.3 Degenerate Cases

Any two homographies $\mathbf{H}_1$ and $\mathbf{H}_2$ induced by a pair of scene planes π_1 and π_2 can be combined as $\mathbf{H} \sim \mathbf{H}_2^{-1}\mathbf{H}_1$, where $\mathbf{H}$ would always be a homology. An intriguing question that may arise is then the following: If this is true for any two scene planes, then why does the similarity measure based on the eigenvalue constraint proposed above work? and when would this constraint degenerate, i.e. fail to determine that the scene triplets undergo the same motion?

To answer these questions, let us re-examine what we do. A homography induced by a scene plane between two views requires a minimum of four points in order to be specified. We only have three points (i.e. the points of a triplet). However, in our case, the fundamental matrix $\mathbf{F}$

is known - we compute it using all the 11 body points across multiple frames. The key idea that makes it possible to compute the homographies is the fact that all the points on the baseline of the two cameras can be transferred via any scene plane. This is because all the points on the baseline are imaged at the two epipoles, and the epipoles can also be considered as the images of the intersection of the scene plane with the baseline. Therefore, when the fundamental matrix is known, one can use the epipoles as the fourth point for the homography induced by any scene plane. Next using the notations of Fig. 3, the homology H maps the points xi as follows:

$$\mathbf{H}\mathbf{x}_i \sim \mathbf{H}_2^{-1}\mathbf{H}_1\mathbf{x}_i. \tag{2.18}$$

Using equation (2.14)

$$\mathbf{H}\mathbf{x}_i \sim \mathbf{H}_2^{-1}\mathbf{y}_i \tag{2.19}$$

By Desargues' theorem [25], lines joining corresponding points must intersect at the vertex of the homology, which in our case is the epipole. This implies that $\mathbf{H}_2^{-1}\mathbf{y}_i$ must lie on the epipolar line of $\mathbf{y}_i$, that is $\mathbf{e}$, $\mathbf{x}_i$ and $\mathbf{H}_2^{-1}\mathbf{y}_i$ must be collinear. This can be expressed as

$$\mathbf{x}_i^T[\mathbf{e}]_\times\mathbf{H}_2^{-1}\mathbf{y}_i = 0. \tag{2.20}$$

A similar result can be established for $\mathbf{H}_1$. This reveals an interesting result: the homographies induced by a moving triplet must be consistent with the fundamental matrix. This constraint implies that the matrix $\mathbf{H}_i^T\mathbf{F}$ is skew-symmetric [29]. Since we use the epipoles in our computation of the homographies, the consistency condition is satisfied when the points of the triplets viewed in both cameras start and end in the same positions up to a similarity. However, this is not the only case where the consistency is preserved, since if the points move in such a way that the vertices of the

triplet remain along the lines joining the second camera center and the vertices of the triplet, the consistency with the fundamental matrix is still preserved. Fortunately, however, for two different actions it is highly unlikely that this can occur to a body triplet, and even less likely to happen for all possible triplets (e.g. to all the 165 triplets specified by the 11 body points).

In addition to the degenerate case, since homography is computed using epipole correspondence, the same consideration of degenerate triplets should be taken in computing the eigenvalue equality invariant.

2.5 Algorithm for Matching Pose Transitions

We now summarize our algorithm for matching two pose transitions $I_1 \rightarrow I_2$ and $J_i \rightarrow J_j$ based on the geometric invariants described above, as follows:

1. Suppose that N body point correspondences are given by $\mathbf{m}_n^1, \mathbf{m}_n^2, \mathbf{m}_n^i, \mathbf{m}_n^j$, for I_1, I_2, J_i and J_j, respectively, where $n = 1, \ldots, N$.

2. Compute the fundamental matrix $\mathbf{F}$, such that $(\mathbf{m}_n^1)^T \mathbf{F} \mathbf{m}_n^i = 0$, $(\mathbf{m}_n^2)^T \mathbf{F} \mathbf{m}_n^j = 0$.

3. Find the epipoles by solving $\mathbf{F}\mathbf{e}_1 = 0, \mathbf{F}^T\mathbf{e}_2 = 0$.

4. For each non-degenerate triplet Δ_i, compute the difference of its motions in $I_{1\rightarrow 2}$ and $J_{i\rightarrow j}$, $E(\Delta_i)$. Based on the selection of geometric invariants, $E(\Delta_i) = E_f(\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2)$ or $E(\Delta_i) = E_h(\mathbf{H})$, where $\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2$ and $\mathbf{H}$ are computed as described in section 2.3 and 2.4.

5. Compute the error function for pose transitions

$$\mathcal{E}(I_1 \to I_2, J_i \to J_j) = \underset{\text{all } \Delta_i}{\text{Median}}(E(\Delta_i)). \tag{2.21}$$

If $\mathcal{E}(I_1 \to I_2, J_i \to J_j) < \tau$, where τ is a small value threshold, then $I_1 \to I_2$ and $J_i \to J_j$ are similar.

2.6 Experimental Results

In this section, we study the properties of our geometric invariants on semi-synthetic data, which are generated from real motion-capture data using synthetic cameras with controlled intrinsic parameters, different viewing directions, and varying noise levels. We generated our data set using the CMU Motion Capture database (MoCap)[3], which consists of sequences of various actions in 3D, captured from real human actions. Here, we do not use the 3D points provided by the data, but merely project the 3D points onto images through synthetic cameras. In other words, we generate the images of 3D body points of a true person, using synthesized cameras and add Gaussian noise. Instead of using all the body points provided in the database, we selected a small subset of body points, which our experiments showed to be sufficient to represent human actions. The body model we employed consists of 11 joints and end points, including head, shoulders, elbows, hands, knees and feet (see Fig. 2.1). Experiments were then carried out on these generated 2D data to evaluate the performance of our method in recognizing pose transitions in the presence of noise, varying viewpoints, and subject-dependent differences.

[3] http://mocap.cs.cmu.edu/

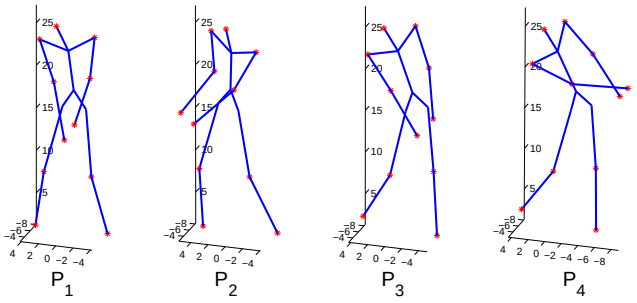


Figure 2.5: Two different pose transitions $P_1 \rightarrow P_2$ and $P_3 \rightarrow P_4$ from a golf swing action.

2.6.1 View Invariance

We selected four different poses P_1, P_2, P_3, P_4 from a golf swinging sequence (see Fig.2.5). To verify the proposition 3, we generated two cameras as shown in Fig.2.6 (a): camera 1 was placed at an arbitrary viewpoint (marked by red color), with focal length $f_1 = 1000$; camera 2 was obtained by rotating camera 1 around an axis on x-z axis plane of camera 1 (colored as green), and a second axis on y-z axis plane of camera 1 (colored as blue), and changing focal length as $f_2 = 1200$. Let I_1 and I_2 be the images of poses P_1 and P_2 on camera 1 and I_3, I_4, I_5 and I_6 the images of poses P_1, P_2, P_3 and P_4 on camera 2, respectively. Two sets of pose similarity errors were computed at all camera positions shown in Fig.2.6 (a): $\mathcal{E}_f(I_1 \rightarrow I_2, I_3 \rightarrow I_4)$ and $\mathcal{E}_f(I_1 \rightarrow I_2, I_5 \rightarrow I_6)$. The results are plotted in Fig.2.6 (b) and (c), which show that, when two cameras are observing the same pose transitions, the error is zero regardless of their different viewpoints, confirming proposition 3.

Similarly, we fixed camera 1 and moved camera 2 on a sphere as shown in Fig.2.6 (d). The errors $\mathcal{E}_f(I_1 \rightarrow I_2, I_3 \rightarrow I_4)$ and $\mathcal{E}_f(I_1 \rightarrow I_2, I_5 \rightarrow I_6)$ are computed and shown in Fig.2.6 (e) and

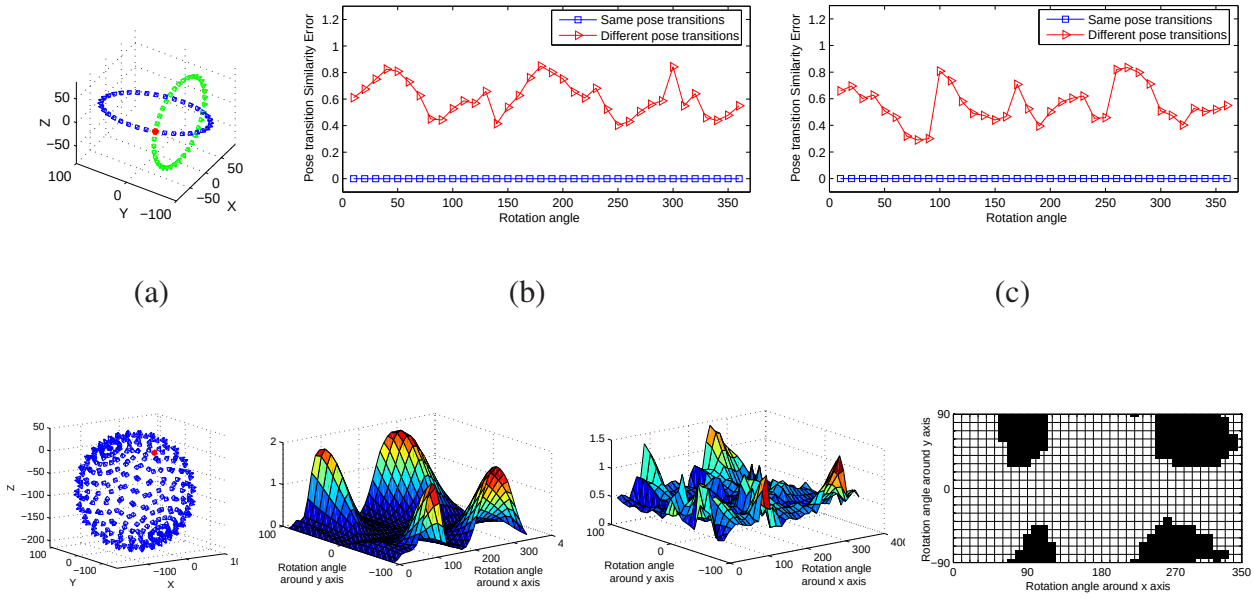


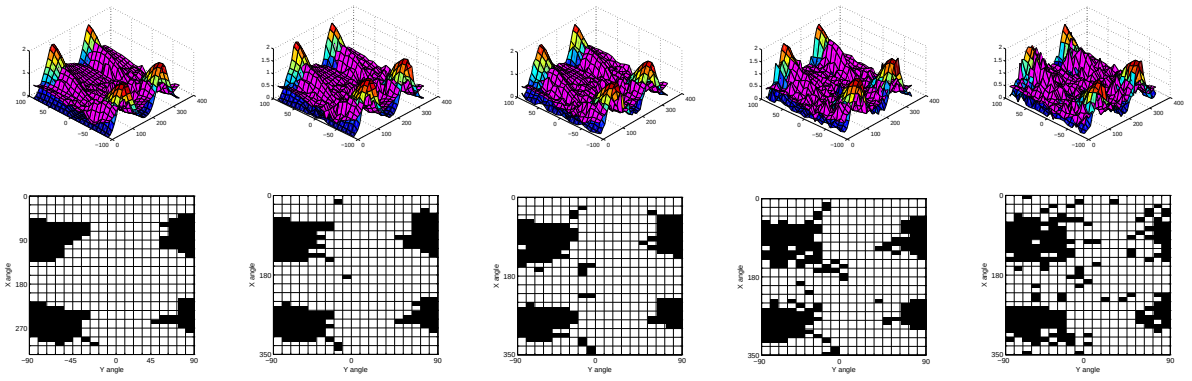
Figure 2.6: Analysis of view invariance. (a) Camera 1 is marked in red, and all positions of camera 2 are marked in blue and green. (b) Errors for same and different pose transitions when camera 2 is located at viewpoints colored as green in (a). (c) Errors of same and different pose transitions when camera 2 is located at viewpoints colored as blue in (a). (d) General camera motion: Camera 1 is marked as red, and camera 2 is distributed on a sphere. (e) Error surface of same pose transitions for all distributions of camera 2 in (d). (f) Error surface of different pose transitions for all distribution of camera 2 in (d). (g) The regions of confusion for (d) marked in black (see text).

(f). Under this more general camera motion, the pose similarity score of corresponding poses is not always zero, since the epipoles in equations (2.5) and (2.6) are approximated. However, this approximation is inconsequential in most situations, because the error surface of different pose transitions is in general above that of corresponding pose transitions, which enables us to readily identify same pose transitions. Fig.2.6 (h) shows the regions (black colored) where approximation is invalid. These regions correspond to the situation that the angles between camera orientations is around 90 degrees, which usually implies severe self-occlusion and lack of corresponding points in practice.

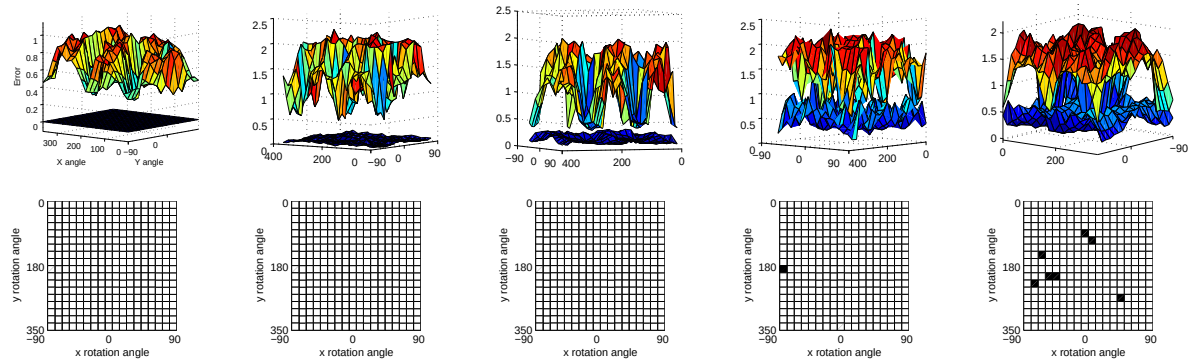
Another two sets of errors $\mathcal{E}_h(I_1 \rightarrow I_2, I_3 \rightarrow I_4)$ and $\mathcal{E}_h(I_1 \rightarrow I_2, I_5 \rightarrow I_6)$ are also computed and shown in figure 2.7(a). The lower flat plane in 2.7(a) demonstrates that when two cameras are observing the same pose transition, the error is always zero in all camera configurations. As shown in this figure, pose transition difference is readily recognized using eigenvalue equality invariant, regardless of the changes in the viewpoint and camera intrinsic parameters.

2.6.2 Robustness to Noise

We used the same data P_1, P_2, P_3, P_4 in section 2.6.1 and cameras configuration in figure 2.6 (d), and then added Gaussian noise to the image points, with σ increasing in steps of 0.25 from 0 to 6.75. Four sets of errors $\mathcal{E}_f(I_1 \rightarrow I_2, I_3 \rightarrow I_4)$, $\mathcal{E}_f(I_1 \rightarrow I_2, I_5 \rightarrow I_6)$, $\mathcal{E}_h(I_1 \rightarrow I_2, I_3 \rightarrow I_4)$ and $\mathcal{E}_h(I_1 \rightarrow I_2, I_5 \rightarrow I_6)$ were computed. For each noise level (σ), the above procedure was run for 50



(a) $\sigma = 0$ (b) $\sigma = 1$ (c) $\sigma = 2$ (d) $\sigma = 3$ (e) $\sigma = 4$



(a) $\sigma = 0$ (b) $\sigma = 1$ (c) $\sigma = 2$ (d) $\sigma = 3$ (e) $\sigma = 4$

Figure 2.7: Robustness to noise: The first row shows the plots of error surfaces under different noise levels, with cameras configuration as in Fig. 2.6 (d). The black blocks in the second row show the camera configurations when there is confusion between same and difference pose transition.

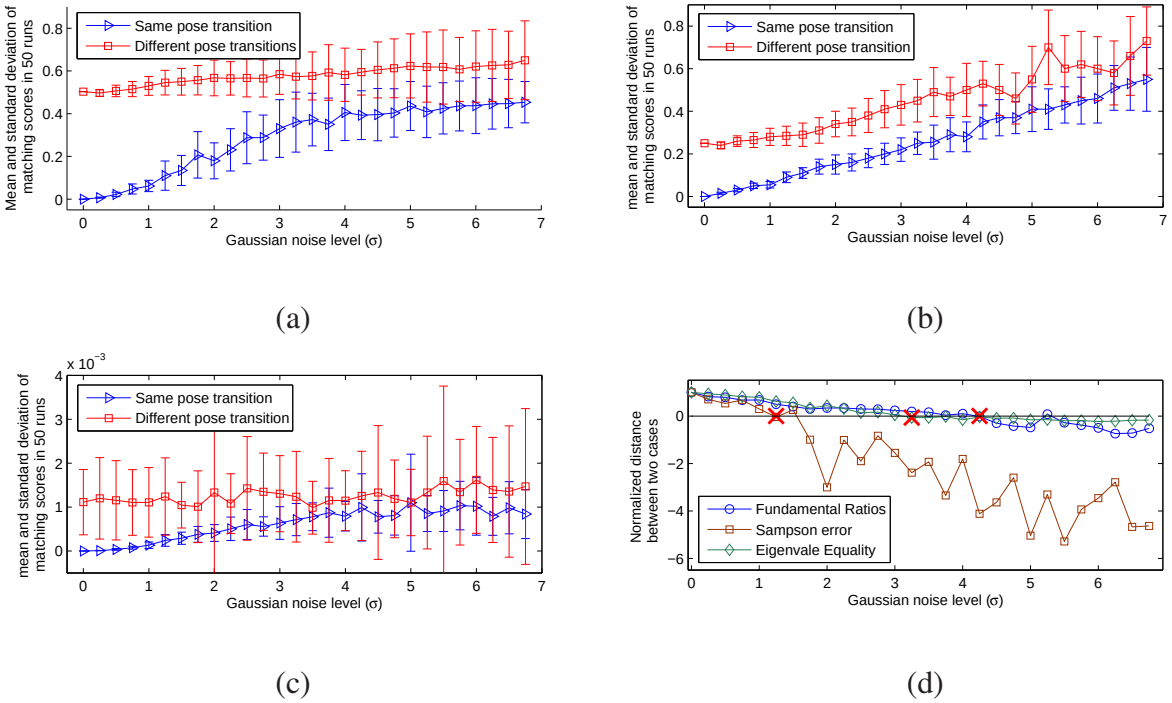


Figure 2.8: Results of using our error functions against the one based on Sampson error: (a) ,(b) and (c) show the plots of matching scores of same and different pose transitions with increasing Gaussian noise for $\mathcal{E}_f(.)$, $\mathcal{E}_h(.)$ and the $\mathcal{E}_s(.)$ (see Appendix A.1), respectively. (d) shows the confusion margin in (a), (b) and (c) (see text).

independent trials and the mean and the standard deviation of all error functions were calculated. The mean error surfaces and confusion areas (black areas) with $\sigma = 0, 1, 2, 3, 4$ are shown in Fig.2.7, in which the (a) - (d) correspond to $\mathcal{E}_f(I_1 \rightarrow I_2, I_3 \rightarrow I_4)$ and $\mathcal{E}_f(I_1 \rightarrow I_2, I_5 \rightarrow I_6)$, and (e) - (h) correspond to $\mathcal{E}_h(I_1 \rightarrow I_2, I_3 \rightarrow I_4)$ and $\mathcal{E}_h(I_1 \rightarrow I_2, I_5 \rightarrow I_6)$. Fig. 2.8 (a) and (b) shows the result of configuration which corresponds to configuration $(0, 90)$ in the second row of Fig. 2.7 (a) - (d), in which case the angle of the camera viewing directions is 90 degrees. Two cases (same and different pose transitions) are readily distinguishable until σ is increased to 4.25, i.e., up to

possibly 12.75 pixel errors. Note that in this experiment the images of the subject have a width of around 150 pixels (see Fig.2.5), which indicates that our method performs extremely well under sever noise.

We compared our results to those obtained using a more classical distance function based on Sampson's error [HZ04] (see Appendix A). The plots are shown in Fig. 2.8 (c). To compare the results in Fig. 2.8 (a) ,(b) and (c), we also computed what we refer to as the *confusion margin* for each likelihood function, which is obtained by computing the distance $d(\sigma)$ between minimum of same pose error bars and maximum of different pose error bars at each noise level σ, and then normalizing it using $\hat{d}(\sigma) = d(\sigma)/d(0)$. If the confusion margin is negative, then the error bars overlap, which indicates confusion in recognizing same and different poses. The curves of both likelihood functions are plotted in Fig. 2.8 (d), and where they go negative are marked by red crosses. Fig. 2.8 (d) shows that our likelihood function is more robust than one based on classical Sampson's error.

2.6.3 Matching Pose Transitions

We selected 40 poses from the MoCap database: 20 from running motion, 10 from golf swing motion, and 10 from walking motion. Each pose is performed by different actors (6 actors for running, 10 for golf swing, and 6 for walking). For each pose, one actor was selected, and the corresponding pose transition was pictured by a camera with arbitrary focal length and viewpoint,

Metric	Using Eigenvalue Equality	Using Fundamental Ratios
Tot. Recognition	11,520	11,520
True Recognition	10,785	10,475
True Recog. %	93.62%	90.93%
Mis-recognition	735	1045
Mis-recognition %	6.38%	9.07%

Table 2.1: Results on testing pose recognition on MoCap data.

to be used as the template for the pose to be recognized. We thus built a database of templates for the 40 pose transitions, denoted as DB. The rest of the pose instances were used to generate testing data by projecting each instance onto images through 48 cameras distributed on a hemi-sphere, as shown in Fig. 2.9. We thus have a test dataset of totally $(5 \times 20 + 9 \times 10 + 5 \times 10) \times 48 = 11,520$ 2D pose transitions, with a vast variety of actor behaviors, camera calibrations and viewpoints. Fig.2.10 shows an example of the same pose observed by some of these cameras. Then each 2D pose transition in the test data is matched against each template in DB, using both $\mathcal{E}_f$ and $\mathcal{E}_h$, and the recognition result is shown in Table 2.1.

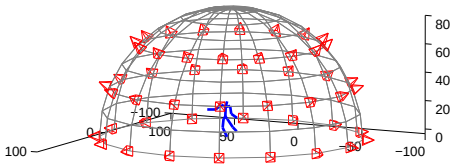


Figure 2.9: The distribution of cameras used to evaluate view-invariance and camera parameter changes in pose recognition using semi-synthetic data.

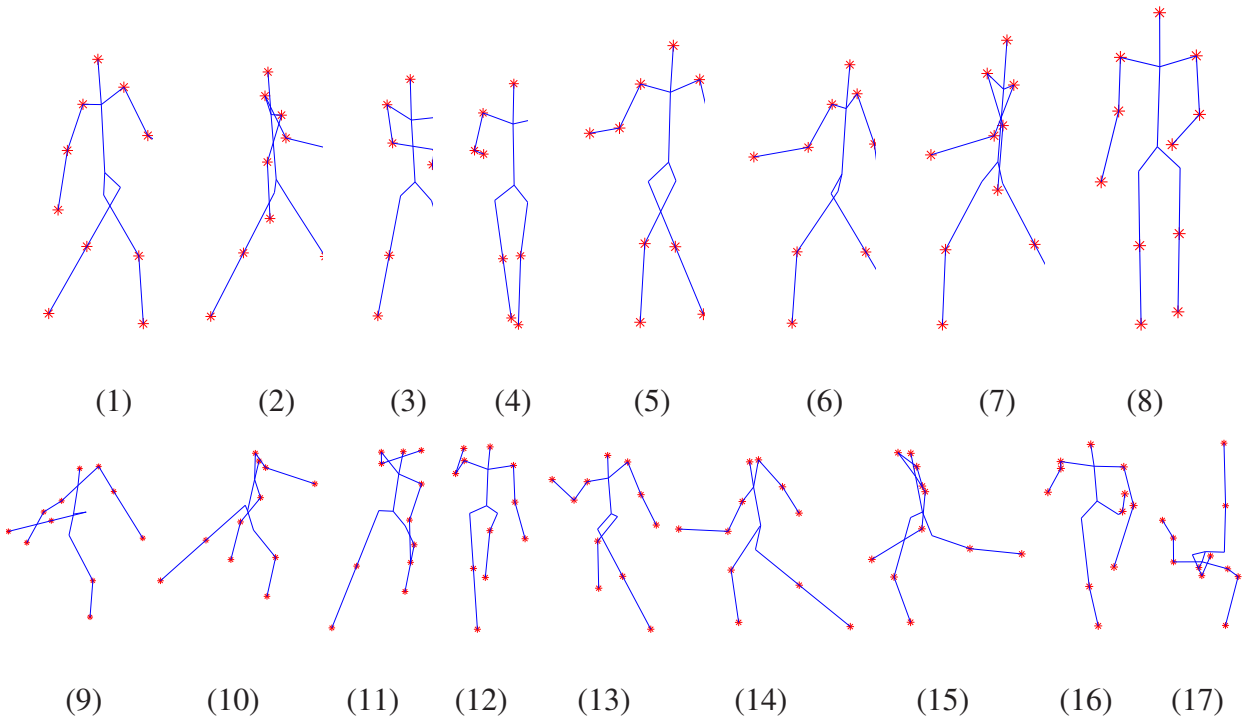


Figure 2.10: A pose observed from 17 viewpoints. Note that only 11 body points in red color are used. The stick shapes are shown here for better illustration of pose configuration and extreme variability being handled by our method.

3 VIEW-INVARIANT RECOGNITION OF HUMAN BODY POSE AND ACTION

As shown in chapter 2, the geometric invariants based on either fundamental ratios or homology eigenvalue equality can capture the trait of non-rigid motion of human body, and are very successful in recognize pose transition from a wide range of viewpoints. In this chapter we demonstrate the application of our geometric invariants to more difficult problems of view-invariant recognition of human pose and action.

3.1 Human Pose Recognition from Video

Pose recognition is an important topic in human motion analysis, and plays a crucial role in many applications such as human motion capture, surveillance systems, HCI, etc. An example of its use in daily life is the interactive karaoke system built by Sun et al. [SLW98], in which human postures are used to trigger and control the karaoke system. In video surveillance systems, the recognition of body poses of monitored people allows for inference of their activities and interactions. Despite many approaches proposed in recent years, the problem remains by and large challenging due to:

(1) large number of degrees of freedom of non-rigid human body [Zat02], which leads to large variations of its appearance. (2) high variability in anthropometry due to a variety of factors such as age, gender, ethnicity, etc. (3) loss of 3D Euclidean information due to perspective camera, including self-occlusions, projective distortions, and dramatic changes in appearance of the same body pose from different viewpoints.

Until recently, research on body pose recognition, such as [ST01, MG00], was primarily focused on spatial information only, i.e. recognition of body pose in a single image. More recent work, has been also investigating temporal information [DLF06], i.e. study human body pose in the context of video data. Our research falls in this latter category. To achieve our goal of recognizing a pre-defined human body pose in a video sequence, we propose a 2D approach exploiting both spatial and temporal information with constraints derived from the multiple view geometry associated with articulated motion.

3.1.1 Related work

There has been substantial work on estimating human pose in 2D images, such as [RF03, RMR04, ZCL04, FH00, RS00b]. These work are primarily focused on tasks such as tracking or detection of body parts or joints, which provide inputs for further recognition of human body poses. Others, as discussed below, focus mainly on the recognition task.

Existing body pose recognition methods can be divided into two main categories: 3D approaches and 2D approaches. The 3D approaches try to recover the 3D body pose from 2D images and compare the body poses in the 3D Euclidean space. These approaches can be further divided into two groups: *model based approaches*, e.g. [DR05, KM00, SBF00, BM98], and *learning based approaches*, e.g. [AT04, SKL05, MM06, OMB06]. The basic idea in model based approach is to approximate the body pose structure by a 3D parametric model, and to estimate its parameters such that the projection of the model closely fits to the human in the images. A key problem in this approach is to define a good likelihood model in terms of edges [DR05, ST01, ST02, Tay00, ST01, ST03, MG00], silhouettes [DR05, KM00], intensities [ST01], or body joints [Tay00]. Another important issue is how to improve the multi-dimensional search in the model parameter space, since it is very expensive. Deutscher et. al. [DR05] proposed an annealing particle filtering system with multiple cameras, which is regarded as an extension of [GD96]. In their system the search for parameters is driven by noise proportional with their individual variances. Another multi-view approach based on contours is reported by Kakadiaris et. al. [KM00]. Sidenbladh [SBF00] use the similar particle filtering technique with statistical human model learned from a discrete set of samples to track people in a monocular sequence. Bregler et. al. [BM98] assumed scaled orthographic projection and they used the product of exponential maps and twist motion to turn the problem as a linear estimation problem. Taylor [Tay00] assumes a scaled orthographic projection model with an infinite number of solutions parameterized by a single scale parameter. By arbitrarily fixing this scale parameter, and hence exploiting the symmetries about a plane parallel to the image plane, he then limits the solution set to a finite

number. The limitations of the method are the arbitrary choice of the scale factor, and the fact that the method would fail in the presence of strong perspective effects. In [ST01] Sminchisescu and Triggs designed a robust likelihood model that combines optical flow, edge energy and motion boundaries. They improved the search efficiency by using a hybrid search algorithm that combines inflated-covariance-scaled sampling and continuous optimization subject to physical constraints. They also used simple kinematic reasoning to enumerate the tree of possible forward/backward flips, and speed the search within each linked group of minima [ST03]. Another work based on MCMC is also reported [MG00]. Moeslund et. al. [MG00] represent the human model in a phase space spanned by its different degrees of freedom and use the analysis-by-synthesis approach to match the model with real images. They further speed their system by pruning the phase space and applying constraints based on human motor system. The learning based approaches avoid the high dimensional search in model parameter space by constraining the possible human poses to a small set, based on the fact that the human poses in ordinary scenes constitute only a small subset of the set of kinematically possible poses. These approaches try to learn a model from training data to recover poses directly from images. To account for camera viewpoint variations, the training data usually consist of several 2D observations of plausible 3D poses. Agarwal and Triggs [AT04] encode the image silhouettes by histogram-of-shape-context descriptors, and learn a compact model from training data by nonlinear regression on the descriptor vectors extracted from training images. Sminchisescu et. al. [SKL05] use discriminative density propagation to estimate 3D motion from image silhouettes. The pose recovery of an image is then performed by simply feeding the image descriptor to the model, which outputs the 3D pose as a 55-dimension vector.

The recognition of given 2D pose data is performed by finding the best match of the 2D observation, and the associated 3D pose is regarded as the reconstruction of a given pose. Mori et. al. [MM06] proposed a 2D exemplar based method based on shape context matching and kinematic chain-based deformation model to estimate 2D body joints on images, and recover the 3D pose using [Tay00]. Ong et. al. [OMB06] presented a exemplar-based 3D human tracking system based on clustering motions ad particle filtering.

Temporal information is also taken into account in some existing methods. Howe et. al. [HLF99] divide 2D tracked human body points in a sequence into "snippets" (11 successive frames) and reconstruct poses per snippet at a time. They model a snippet as a mixture of Gaussian probabilities in a high-dimensional space, and reconstruct it by finding the best maximum *a posteriori* 3D snippet for each of the 2D observations (2D tracked points). Dimitrijevic et. al. [DLF06] proposed a contour based method, in which they synthesize multi-view silhouettes of 3D poses from motion capture data, and for each pose three successive frames are stacked as a temporal template. The edge images of given frames are then matched against all multi-view temporal templates in the database and the best matched template gives the estimated 3D pose. There are also a few *2D approaches* that do not require a 3D parametric body model or the prior knowledge of 3D configurations of body poses. These approaches represent human body pose by 2D features from the images. Bradski et. al. [BD02] propose a method using timed motion history image (tMHI) to recognize human pose in a sequence, which is an extension of work by Bobick and Davis [BD01b]. However, these works do not take into consideration the variations in camera viewpoint and intrinsic parameters, thus are limited in practice under wide camera variability.

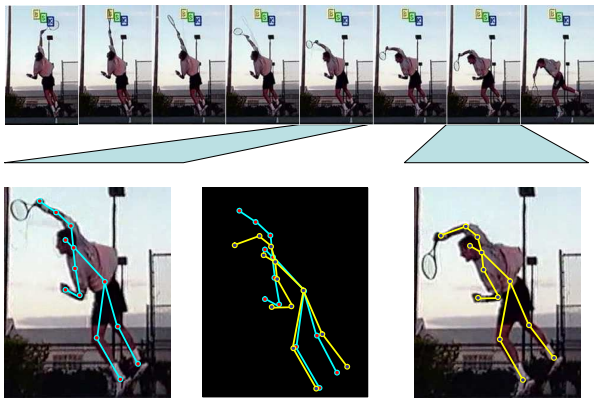
3.1.2 Overview of our pose recognition system

For clarity, here we describe the problem of pose recognition discussed in this thesis:

Pose recognition in a video sequence: Given a video sequence of human motion, which is represented by a sequence of 2D body poses $\{I_{1...n}\}$, our goal is to recognize a set of predefined poses $\mathcal{P}$ in the sequence. The video sequence can be from an uncalibrated camera and an unknown viewpoint.

There are mainly two schools of research on pose recognition: the first, such as [DLF06, HLF99], recognizes human pose directly from low level features on the image, such as edges, silhouettes, optical flow, etc., while the second school such as [MM06, Tay00] uses a detection-recognition scheme: first detect and identify body primitives (e.g. joints or body parts), then recognize the relative position of the detected primitives as that of a known body pose. We follow the latter approach, with the focus in this thesis on the recognition stage. The detection and tracking problem is beyond the scope of this thesis and has been widely studied in decades, e.g., [RK95, PRC99, RF03]. In this thesis, we assume that tracking has been performed, and we are given labelled body points (see Fig.2.1) on video frames as inputs to our system.

We represent the body pose by a space-time template: Spatial information is represented by a set of 2D imaged body points, and the temporal information by the transition between two poses. Unlike temporal templates proposed in [DLF06] or snippets in [HLF99] that require to be generated from multiple viewpoints, our template is generated from only a single arbitrary viewpoint, and the recognition is achieved using geometric constraints imposed by the motions of body point triplets,



(a) (b) (c)

Figure 3.1: An example of a space-time template composed of two poses: (a) the key pose and (c) the succeeding pose. The two poses are overlapped in (b) to show their differences.

which we show to be invariant to camera calibration matrix and changes in viewpoints. View-invariance in most exemplar-based approaches is heavily dependent on the number of viewpoints used in the training data: using a large number of viewpoints will dramatically increase the time complexity, while using a small number may cause poor recognition. Our solution eliminates these problems since only one viewpoint is required and the constructed likelihood function is independent of camera viewpoints and its intrinsic parameters.

3.1.3 Pose Recognition Using Spatial-temporal Template

To recognize body poses, we maintain a set of space-time templates for a selected set of poses to be recognized. For each body pose, we require only one template from any arbitrary viewing

angle. A body pose is recognized by matching the input video against all available templates and choosing the one with highest score, provided that the likelihood is above some threshold τ.

Our space-time templates require only 2D information (2D image coordinates of body joints), and can be either extracted from a real video sequence or synthesized from motion capture data. A template is composed of a pair of 2D poses in order: a key pose which represents the specific pose of interest and a succeeding one, which captures the transition shortly after the key pose. The succeeding pose can be selected arbitrarily, as long as it is sufficiently distinct from the key pose. Fig. 3.1 shows an example of a space-time template extracted from a tennis-serve video sequence.

Unlike most exemplar based approaches that require a large number of training views to achieve reliable view-invariant recognition, we require only one viewing direction for a space-time template. The view invariance in our approach is achieved by deriving a likelihood function $\mathcal{L}(.)$ based on the error function for matching pose transitions in equation 2.9.

Here we describe our solution for recognizing human poses in a video sequence taken by an uncalibrated camera from arbitrary viewpoints. Suppose we are given a video sequence $\{I_{1\ldots n}\}$, where I_i is the body point representation of the human pose in i^{th} frame, and n is the total number of frames. We are also given the temporal template $T^k_{1,2}$ of some known pose $\mathcal{P}^k$. The procedure for recognizing $\mathcal{P}^k$, if it exists in $\{I_{1\ldots n}\}$, is described as follows:

1. For each frame I_i, $i = 1 \ldots n$, find its d succeeding frames in the video $\{I_{i+1}, I_{i+2}, \ldots, I_{i+d}\}$.

2. The body pose in I_i is recognized as $\mathcal{P}^k$ if

$$\max_j \left\{ \mathcal{L}(I_{i,j}, T^k_{1,2}) | j = i+1, \ldots, i+d \right\} > \tau,$$

where τ is a threshold and

$$\mathcal{L}(I_{i,j}, T^k_{1,2}) = 1 - \mathcal{E}(I_i \rightarrow I_j, T^k_1 \rightarrow T^k_2). \tag{3.1}$$

Checking a segment of d frames allows our method to accommodate for different frame rates of videos and varying execution rates of the motions. As shown in section 3.2.4, our method can recognize poses from real video data, when camera calibration is unknown and viewpoints are extremely different.

3.2 View Invariant Action Recognition

Action recognition could be regarded as an extension of the pose recognition problem, added the dynamics of human body. It is a challenging problem that combines the uncertainty associated with computational vision and unpredictable human behavior. In addition to those factors discussed in 3.1, including large number of degrees of freedom of non-rigid human body, high variability in anthropometry, and viewpoint changes, the temporal variability of action sequences is another source of challenges in action recognition. The execution rates of the same action in different videos may vary due to different actors or variable camera frame rates. Therefore, the mapping between same actions in different videos is usually highly non-linear.

To make the problem more tractable, most researchers have made simplifying assumptions on two or more of the following aspects:

- Camera model : such as scaled orthographic e.g. [SS05], or assumed a calibrated camera in the case of a perspective camera.
- Camera pose : i.e. little or no viewpoint variations.
- Anatomy : such as isometry [PC03], coplanarity of a subset of body points [PC03], etc.

Space-time features are essentially the primitives that are used for recognizing actions, e.g. photometric features such as the optical flow [EBM03, ZXG06, Wan06] and the local space-time features [SLC04, LBP05]. These photometric features can be affected by luminance variations due to, for instance, camera zoom or pose changes, and often work better when the motion is small or incremental. On the other hand, salient geometric features such as silhouettes [BGS05, WS07, BD01b, WTN03, YS05] and point sets [PC03] are less sensitive to photometric variations, but require reliable tracking. Silhouettes are usually stacked in time as 2D object, such as temporal template [BD01b], or 3D objects [BGS05, YS05], while point sets are tracked in time to form a set of space-time curves.

Some existing approaches are also more holistic and rely on machine learning techniques, e.g. Hidden Markov Model (HMM) [AL06], Support Vector Machine (SVM) [SLC04], etc. As in most exemplar-based methods, these techniques rely on the completeness of the learning data, and are usually expensive when it is required to learn a model from a large dataset.

3.2.1 Related Work

Most action recognition methods adopt simplified camera models and assume that camera viewpoint is fixed or simply ignore the effect of camera viewpoint changes on action recognition. However, in practical applications such as surveillance, actions may be viewed from different angles by different perspective cameras. Therefore, a reliable action recognition system has to be invariant to the camera viewpoint or its parameters. View-invariance is, thus, of great importance in action recognition, and has received relatively more attention in recent literature.

One approach to tackle view-invariant pose and action recognition has been based on using multiple cameras: Campbell et al. [CBA96] use stereo images to recover a 3D Euclidean model of the human subject, and extract view invariance for 3D gesture recognition; Weinland et al. [WRB06] use multiple calibrated and background-subtracted cameras to achieve the goal of view-invariant action recognition. They obtain a visual hull for each pose from silhouettes extracted in multiple views, and stack them to generate a motion history volume, based on which Fourier descriptors are computed as the representations of actions. Ahmad and Lee [AL06] proposed an HMM-based method using multiple cameras. They build HMMs on optical flow and human body shape features in different views, and feed a test video sequence taken from an arbitrary viewpoint to all learned HMMs. A maximum likelihood scheme is then applied to classify the test video. These methods require the setup of multiple cameras, which is quite expensive and restricted in many situations such as online video broadcast or monocular surveillance.

A second line of research is based on a single camera and is motivated by the idea of exploiting the invariants associated with a given camera projection model, e.g. affine, or projective. For instance, Rao et al. [RYS02] propose a view invariant method for recognizing human hand actions. They assume an affine camera model, and use dynamic instants, which are defined as the maxima in the spatio-temporal curvature of the hand trajectory, to characterize hand actions. The difficulties with this representation is that dynamic instants may not always exist, e.g. curvature is constant in a circular trajectory, or may not be always preserved from 3D to 2D due to perspective effects. Moreover the affine camera model is restrictive in most practical scenarios. A more recent work reported by Parameswaran and Chellappa [PC03] relaxes the restrictions on the camera model. They propose a quasi-view-invariant 2D approach for human action representation and recognition, which relies on the number of invariants in a given configuration of a set of points. Thus a set of projective invariants are extracted from the images and used as action representation. However, in order to make the problem tractable under variable dynamics of actions they introduced heuristics, and make simplifying assumptions such as isometry about human body parts. Moreover, they require that at least five body points form a 3D plane during the course of an action.

Another promising approach is based on exploiting the epipolar geometry. Two subjects in the same exact body posture viewed by two different cameras at different viewing angles can be regarded as related by the epipolar geometry. Therefore, corresponding poses in two videos of actions are constrained by the associated fundamental matrices, providing thus a way to match poses and actions in different views. The use of fundamental matrix in view invariant action recognition is first reported by Syeda-Mahmood et al. [SVS01] and later by Yilmaz et al. [YS05, YS06]. They

stack silhouettes of input videos into space-time objects, and extract features in different ways, which are then used to compute a matching score based on the fundamental matrices. A similar work is also presented in [GSS04], which is based on body points instead of silhouettes. The approach presented in this thesis falls in the second category and is based on geometry. We assume a fully projective camera with no restrictions on pose and viewing angles. Moreover, our formulation relaxes restrictive anthropometric assumptions such as isometry. This is due to the fact that unlike existing methods that regard an action as a whole, or as a sequence of individual poses, we represent an action as a set of non-rigid *pose transitions* defined by triplets of points - that is we break down further each pose into a set of point-triplets and find invariants for the motion of these triplets across two frames. Therefore, the matching score in our method is based on *pose transition*, instead of being based directly on individual poses or on the entire action. Our approach can also accommodate the possibility of self-occlusion, which may occur under some poses.

3.2.2 Representation of Action

Since action can be regarded as a sequence of poses, a straightforward approach to match two actions is to check the pose-to-pose correspondences. Two same body poses observed by different cameras are related by epipolar geometry via the fundamental matrix, which provides a cue to match the two poses, regardless of camera calibration matrices or viewpoints. This has motivated the research reported in [GSS04, SS05, SVS01] that are based on the fundamental matrix. Pose-to-pose correspondence is a necessary, but usually not sufficient condition for action correspondence.

(a) (b)

Figure 3.2: Two distinct actions with corresponding poses. (a) The subject keeps the same pose in the sequence. (b) The subject is performing a rotation around an axis.

Consider the following case: A subject keeps the pose as illustrated in Fig. 3.2 (a) during the sequence 1, while in sequence 2 (Fig. 3.2 (b)) it performs a spin, i.e. a rotation around some axis while keeping the same pose as the subject in Fig. 3.2 (a). These two actions are obviously distinct; however there exist many pose-to-pose correspondences between them since the pose remains unchanged during the two actions. Therefore, additional constraints other than pose correspondence are required to tackle this problem. Most fundamental matrix based methods enforce the constraint that all pose-to-pose correspondences share the same epipolar geometry, i.e., the same fundamental matrix, which is critical to the success of these methods. In order to accommodate for camera motion, instead of using a common fundamental matrix for all pose-to-pose correspondences, Yilmaz et. al. [YS06] proposed to estimate a temporal fundamental matrix, which is equivalent to a time sequence of fundamental matrices corresponding to matching poses. However, their method may misclassify the two actions in Fig. 3.2 as the same action, since there exist a number of temporal fundamental matrices that relate these distinct actions.

A limit of fundamental matrix based methods is that they require at least 7 or 8 point correspondences for each pair of poses to measure their similarity. However, in practice, in order to

overcome errors, they require far more points, which may not be always possible, especially when self-occlusions exist. For pose-to-pose based methods, this requirement is repeated for every pair of poses (i.e. every image pair), increasing thus their noise sensitivity. We overcome this problem by breaking down body pose into point triplets leading to a largely over-determined problem as described below.

Since actions are spatio-temporal data in 4D, the temporal information is essential to the perception and understanding of actions. However, this is practically ignored when working directly on individual poses. In this thesis, we regard action as a sequence of pose transitions. In the example shown in Fig. 3.2, although sequences (a) and (b) have the same poses, they are performing different sequences of pose transitions, allowing for distinguishing between the two actions.

Proposition 5 *Two actions are identical if and only if they start at the same pose, and follow the same sequences of pose transitions.*

This proposition implies that the recognition and matching of two actions can be achieved by measuring the similarity between their sequences of pose transitions. The problem is then reduced to matching pose transitions, which is stated as follows: given two pairs of poses, denoted by $\langle I_1, I_2 \rangle$ and $\langle J_i, J_j \rangle$ (see Fig. 2.2), we are to determine whether the transformation from I_1 to I_2 matches to that from J_i to J_j. Note that $I_{1,2}$ and $J_{i,j}$ are sets of 2D labelled points that are observed by cameras with different intrinsic parameters and from different viewpoints.

3.2.3 Action Alignment and Recognition

The goal of action alignment is to determine the correspondences between two video sequences $A = \{I_{1...n}\}$ and $B = \{J_{1...m}\}$ with matching actions, in our case based on pose transitions. We align A and B by seeking the optimal mapping $\psi : A \rightarrow B$ such that the cumulative similarity score $\sum_{i=1}^{n} S(i, \psi(i))$ is maximized, where $S(.)$ is the similarity of two poses. This is solved by Dynamic Programming (DP), which has been proven successful in sequence alignment, and has been applied in many areas, such as text processing, bioinformatics. Its application in action recognition can also be found in [PC03, RYS02]. The key is to define $S(.)$ based on matching pose transitions:

$$S(i, j) = \tau - \mathcal{E}(I_{i \rightarrow r_1}, J_{j \rightarrow r_2}), \tag{3.2}$$

where τ is a constant threshold, and $r_1, s_1 \in [1, n]$ and $r_2, s_2 \in [1, m]$ are computed as

$$\langle r_1, r_2 \rangle = \underset{r_1, r_2}{\operatorname{argmin}} \{ \min_{s_1, s_2} \mathcal{E}(I_{r_1 \rightarrow s_1}, J_{r_2 \rightarrow s_2}) \}. \tag{3.3}$$

The matching score of A and B is then defined by

$$\mathscr{S}(A, B) = \max_{\psi} \sum_{i=1}^{n} S(i, \psi(i)). \tag{3.4}$$

In other words, a pair of reference poses $\langle r_1, r_2 \rangle$ are found first, then two pose-transition sequences $A(r_1)$ and $B(r_2)$ are aligned using DP. The initialization $\langle r_1, r_2 \rangle$ can be further simplified by fixing r_1 and s_1, e.g., $r_1 = \lfloor \frac{1}{4} n \rfloor$ and $s_1 = \lfloor \frac{3}{4} n \rfloor$.

The traced-back path P provides an alignment between two video sequence. Note that this may not be a one-to-one mapping, since there exists horizontal or vertical lines in the path (see Fig. 3.6

(b) for example). In addition, due to noise and computational error, different initializations may lead to slightly different valid alignment results. Here the action matching score rather than the alignment is what we are concerned with in action recognition.

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

3.2.4 Experimental Results

3.2.4.1 Motion Capture Data

We selected 5 actions from CMU's MoCap dataset: walk, jump, golf swing, run, and climb. Each action is performed by 3 actors, and each instance of 3D action is observed by 17 cameras: the first camera was placed on $(x_0, 0, 0)$, looking at the origin of the world coordinate system, while the rest 16 cameras were generated by rotating around the y-axis by β and around the x-axis by α, where $\beta = i\frac{\pi}{4}, i = 0, \ldots, 7$ and $\alpha = j\frac{\pi}{4}, j = 0, 1, 2$ (see Fig. 3.3 for the location of cameras). The focal lengths were also changed randomly in the range of 1000 ± 300.

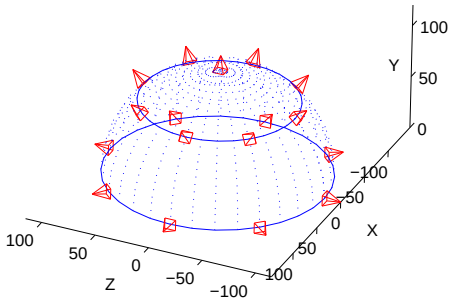


Figure 3.3: The distribution of cameras used to generate action database.

Our dataset consists of totally 255 video sequences, from which we generated a reference Action Database (ADB) of 5 sequences, one sequence for each action. These sequences are all selected from viewpoint 1. The rest of the dataset was used as test data, and each sequence was matched against all actions in the ADB and classified as the one with highest score. For each sequence matching, 10 random initialization are tested. The classification results are shown in Table 3.1 for using eigenvalue equality invariant, and 3.2 for using fundamental ratios invariant. The overall classification accuracy is about 92% in Table 3.1 and 82% in Table 3.2.

A further examination of the results on viewpoints is shown in Table 3.3 for using eigenvalue equality invariant and Table 3.4 for using fundamental ratios invariant. From the results we find that the accuracy for viewpoint 17 is low using both invariants (46.7% and 60.0%). This is most possibly due to the severe distortion of subjects when looking down from viewpoint 17, which is directly above the subject, which is also reported by [PC03]. Ignoring this highly unlikely viewpoint, the average accuracy for other viewpoints is about 95% when using the eigenvalue equality invariant, which is quite outstanding, given the extreme viewpoint changes and camera intrinsic parameter variations. When fundamental ratios invariant is used, the overall classification accuracy

Table 3.1: Confusion matrix using eigenvalue equality invariant: Large values on the diagonal entries indicate accuracy.

Ground-truth	Recognized as				
	Walk	Jump	Golf Swing	Run	Climb
Walk	45	1		2	2
Jump	2	47		1	
Golf Swing	1		48	1	
Run		3		47	
Climb	6	2			42

for all viewpoints is 81:60%, with very low accuracy at viewpoints 11, 14, 15, 16, which correspond to severe viewing angles from below or above the actor. This is consistent with observations pointed out in section 2.6.1. Excluding these viewpoints, the classification accuracy increases to 94:21%.

3.2.4.2 Real Video Data

We also evaluated our method on a dataset of real video sequences. To best simulate the situations in real life, we collected these videos from Internet, coming from a variety of sources. The collected dataset consists of 56 sequences of 8 actions (see Fig. 3.4 for snapshots): 4 of ballet fouette, 12 of ballet spin, 6 of push-up exercise, 8 for golf swing, 4 of one-handed tennis backhand stroke,

Table 3.2: Confusion matrix using fundamental ratios invariant: Large values on the diagonal entries indicate accuracy.

Ground-truth	Recognized as				
	Walk	Jump	Golf Swing	Run	Climb
Walk	39	3	1	5	2
Jump	4	44		1	1
Golf Swing	1	1	45	2	1
Run	4	3		41	2
Climb	8	3	1	3	35

Table 3.3: Recognition accuracy for various viewpoints: using eigenvalue equality invariant

	Viewpoints																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
# of sequences	10	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
# of errors	0	0	1	1	0	1	2	0	2	1	1	0	2	1	1	0	8
Accuracy	1.0	1.0	.933	.933	1.0	.933	.867	1.0	.867	.933	.933	1.0	.933	.867	.933	1.0	.467

Table 3.4: Recognition accuracy for various viewpoints: using fundamental ratios invariant

	Viewpoints																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
# of sequences	10	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
# of errors	0	1	1	0	1	1	1	2	0	1	11	2	0	4	8	7	6
Accuracy	1.0	.933	.933	1.0	.933	.933	.933	.867	1.0	.933	.267	.867	1.0	.733	.467	.533	.600

8 of two-handed tennis backhand stroke, 4 of tennis forehand stroke, and 10 of tennis serve. Each action is performed by different subjects, and the videos are taken by different unknown cameras from various viewpoints collected over Internet. In addition, videos in the same group (action) may have different starting and ending times, thus may be only partially overlapped. Subjects also perform the same action in different ways and at different speeds. We used the human model in Fig. 2.1, which consists of 11 body points. Our experiments show that these 11 points are sufficient for recognition of human action in practically any situation. Self-occlusion also exists in many of the sequences, e.g., the right shoulder points in Fig.3.4 (6), (9), (19), (21), (23), (47), (49) are occluded. Minor motion of camera also exists in some sequences, which is a good test of robustness of our method.

3.2.4.3 Action Recognition in Videos

We selected a number of poses in each action in our real video database, and built templates for them from one of the instances in that group. When building templates, we set the distance between key pose and successive pose as 5 frames, and during recognition phase we tested using $d = 10$ successive frames for each frame in the video.

Examples of templates and recognition results are shown in Fig. 3.5. Three selected poses and their associated templates are shown in Fig. 3.5 (a), (c) and (e), where the key poses of the templates are marked in blue and the succeeding poses are marked in red. (a) and (c) are built from a tennis backhand stroke sequence, and their corresponding poses are recognized in two video

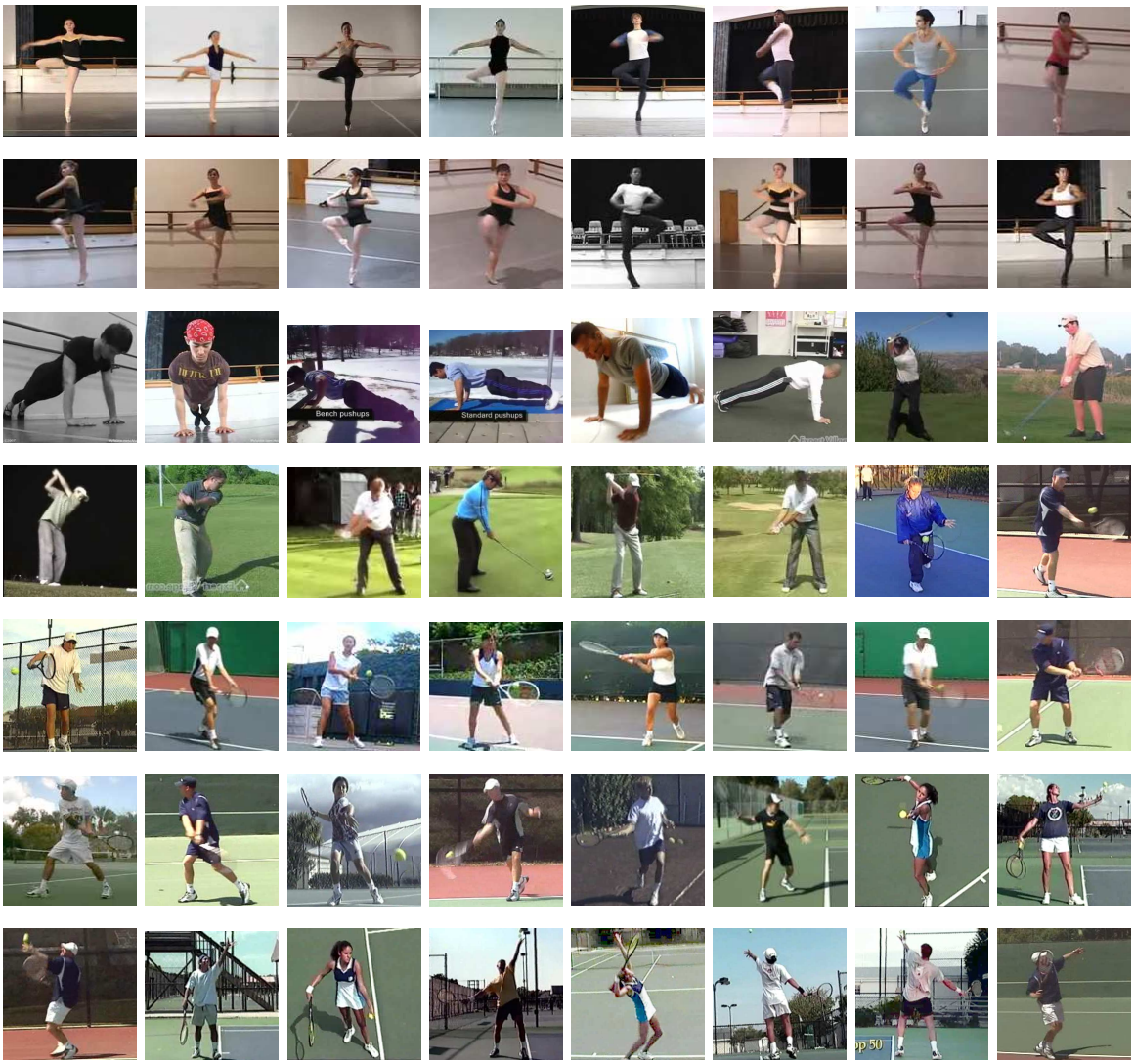
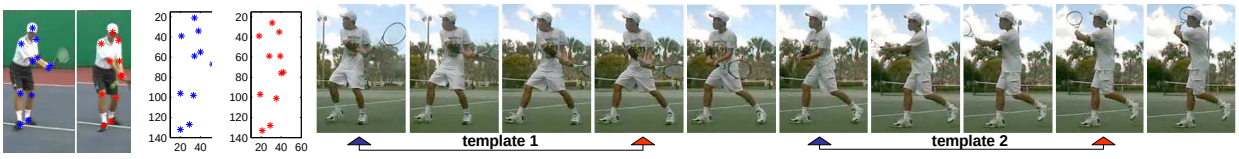
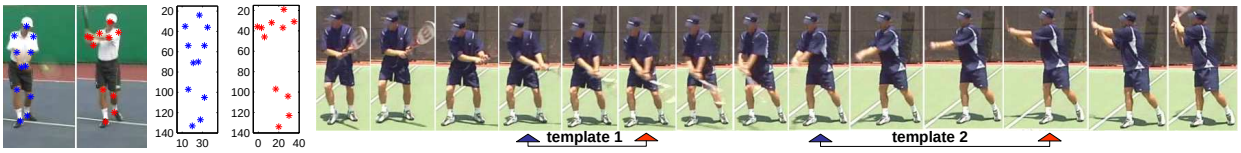


Figure 3.4: A set of 56 sequences in 8 categories (actions) used to test the proposed method. Ballet fouette: (1)-(4); ballet spin: (5)-(16); push-up: (17)-(22); golf swing: (23)-(30); one-handed tennis backhand stroke: (31)-(34); two-handed tennis backhand stroke: (35)-(42); tennis forehand stroke: (43)-(46); tennis serve: (47)-(56).

sequences taken by unknown cameras from different viewpoints. Blue arrows in Fig. 3.5 (b), (d) and (f) indicate the locations of recognized poses (key poses) and red arrows indicate their succeeding poses. Results show that by using a single template from an arbitrary viewpoint, our method can recognize poses from videos captured by unknown cameras with different internal parameters and viewpoints.



(a) template 1 (b) detection of poses associated to template 1 and template 2



(c) template 2 (d) detection of poses associated to template 1 and template 2

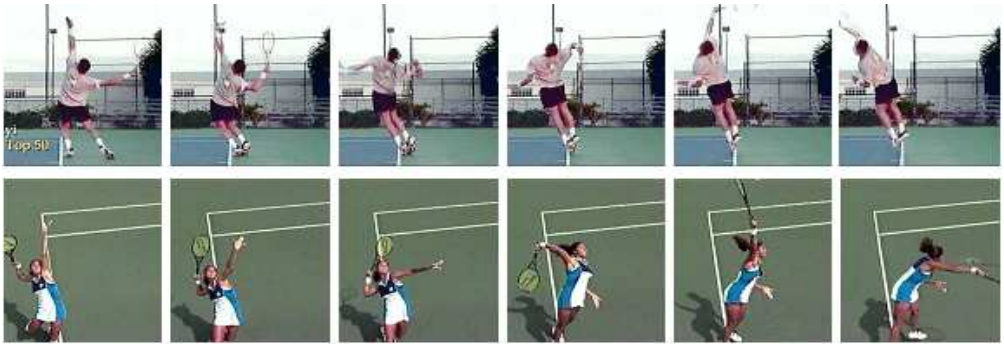
(e) template 3 (f) detection of poses associated to template 3

Figure 3.5: Results of recognizing human poses in video sequences.

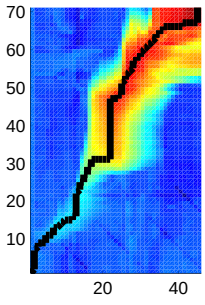
3.2.4.4 Action Sequence Alignment

We tested our action alignment approach for numerous sequences in our database, two of which are shown in Fig. 3.6. These test sequences had different lengths or different starting and ending points

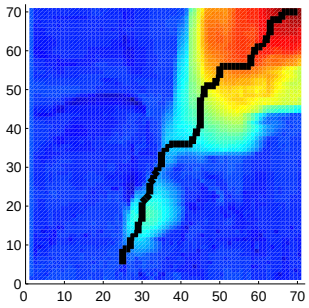
(a) An alignment of two golf swing sequences.



(b) An alignment of two tennis-serve sequences.



(c) $\tau = 0.3$



(d) $\tau = 0.4$

Figure 3.6: Two examples of action alignment: (a) shows the frame-by-frame mappings between the two golf-swing sequences with different lengths, (b) alignment for a tennis-serve action with different starting and ending frames, (c) and (d) show the optimized traced paths using dynamic time warping.

of action. Fig. 3.6 (a) and (b) show the two examples of aligned sequences. In the first example, two sequences of different lengths (the length of the upper sequence is 46 frames, and the lower one is 71 frames) are aligned, in which two players are performing golf swing at different speeds. The alignment result is shown in Fig. 3.6 (a): the first two rows show some matched poses, and the frame-to-frame mapping of two sequences are displayed in the third row. In the second example, shown in Fig. 3.6 (b), two sequences of a tennis serve-actions are aligned: the two sequences are roughly of the same length but different start and ending frames in terms of player's pose.

The accumulation matrices and the back-tracked paths in dynamic time warping for these two examples are shown in Fig. 3.6 (c) and (d), respectively. The thresholds used in these examples were $\tau = 0.3$ and $\tau = 0.4$. The choice of τ reflects our estimate of the influence of matching errors. Studying the optimal choice of τ would require extensive future investigation beyond the scope of this thesis, which may require studying the methods of cross validation and estimation of uncertainty using sequences of covariance matrices. Although, a too small value of τ may cause sensitivity to noise, and too big of values may lead to false alignment, dynamic time warping in general performs well for a large range of values and provides good solutions.

3.2.4.5 Action Recognition

We built an action database ADB by selecting one sequence for each action from Fig.3.4. The other sequences were used as test sequences, and were matched against all actions in ADB. The recognition result is based on the highest matching score for each sequence. We show the confu-

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

Table 3.5: Confusion matrix for using fundamental ratios invariant. The actions are denoted by numbers: 1 - ballet fouette, 2 - ballet spin, 3 - pushup, 4 - golf swing, 5 - one handed tennis backhand, 6 - two handed tennis backhand, 7 - tennis forehand, 8 - tennis serve. The diagonal nature of the matrix indicates high accuracy.

sion matrix for using fundamental ratios invariant in Table 3.5, which indicates an overall 95.83% classification accuracy for real data. A higher accuracy of 100% is achieved using eigenvalue equality invariant, as shown in Table 3.6. As demonstrated in the results, our method provides a successful recognition of various actions by different subjects, regardless of large differences in unknown camera intrinsic parameters and viewpoints. Our method also accommodates substantial self-occlusions and minor camera motions.

Ground-true	Recognized as action							
actions	#1	#2	#3	#4	#5	#6	#7	#8
#1	4							
#2		11						
#3			5					
#4				7				
#5					3			
#6						7		
#7							3	
#8								9

Table 3.6: Confusion matrix for using eigenvalue equality invariant. The labels of actions are same as in Tab. 3.5.

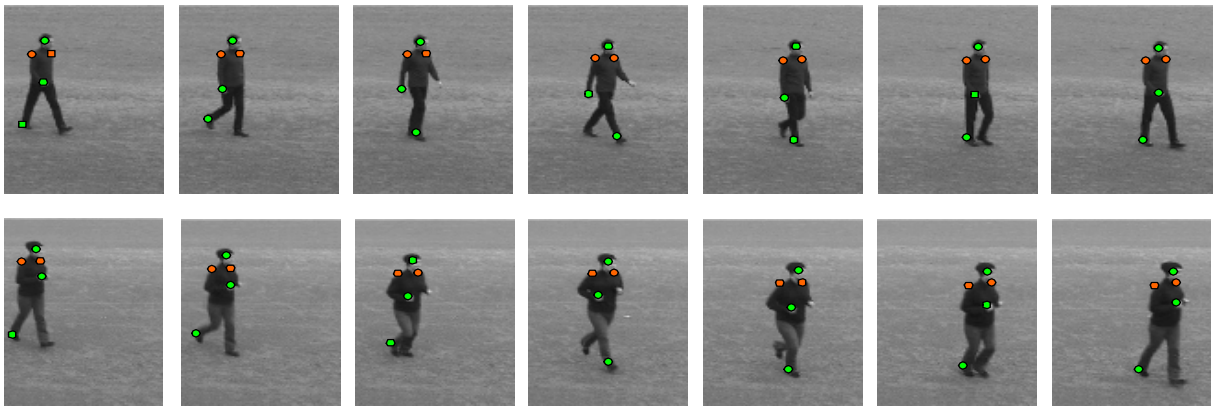
4 WEIGHTING-BASED HUMAN ACTION RECOGNITION

In this chapter, we propose a technique to enhance the application of geometric invariants described in chapter 3 in human motion analysis, by analyzing the contributions of different body point triplets.

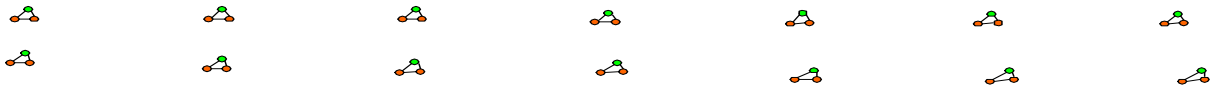
4.1 Introduction

In Chapter 2 and 3, we implicitly made an assumption that all body joints have equivalent contribution to the tasks of matching pose transitions, and the recognition of pose and actions, which is usually not in line with the reality. For example, in some sport games such as boxing, the motion of the upper body parts of the fighter attracts much more attention of the observer than the legs and feet. A person walking with a suitcase may have different motion of the arms and hands from regular walking, while people do not distinguish them as different actions compared with other actions such as running or jumping. Humans tend to pay more attention to body parts that capture the features essential to specific recognition tasks. Using a point-based model of the human body, it is believed that some body points are more important than others to capture the trait of specific actions, e.g., the shoulders, elbows and hands for boxing, and the feet for walking and running.

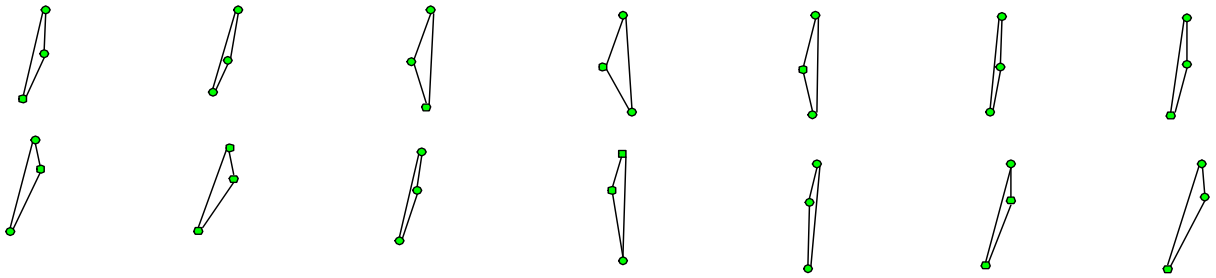
With the triplet representation of human pose and action, as we proposed in chapters 2 and 3, the similar conclusion can be made on body point triplets: some body point triplets have greater contribution to the task of pose and action recognition. For instance, the triplets composed of shoulders and hips have similar motion in walking and jogging, and thus make trivial contribution to distinguish them, while other triplets that consist of shoulder, knee and foot joints carry more essential information of the differences between walking and jogging (See Figure 4.1).



(a) Examples of walking (upper sequence) and jogging (lower sequence)



(b) Comparison of head-shoulders triplet in walking and jogging



(c) Comparison of head-hand-knee triplet in walking and jogging

Figure 4.1: Roles of different triplets in action recognition

Understanding the roles of body point triplets in human motion/action could help us retrieve more accurate information of human motion, and thus improve the efficiency of our system of recognizing human actions.

4.2 Weighting-based Human Action Recognition

To study the roles of body-point triplets in action recognition, we select two different sequences of walking action $WA = \{I_{1...l}\}$ and $WB = \{J_{1...m}\}$, and a sequence of running action $R = \{K_{1...n}\}$. We then align sequence WB and R to WA, using the alignment method described in chapter 3, and obtain the corresponding alignment/mapping $\psi : WA \rightarrow WB$ and $\psi' : WA \rightarrow R$. As discussed in section 3.2.3, the similarity of two poses is computed based on error scores of all body-point triplets motion. For each matched poses $\langle I_i, J_{\psi(i)} \rangle$, we stack the error scores of all triplets as a vector $\mathbf{V}_e(i)$:

$$\mathbf{V}_e(i) = \begin{bmatrix} E(\Delta_1) \\ E(\Delta_2) \\ \vdots \\ E(\Delta_T) \end{bmatrix}, \tag{4.1}$$

and then build an error score matrix $\mathbf{M}_e$ for alignment $\psi_{WA \rightarrow WB}$:

$$\mathbf{M}_e = \begin{bmatrix} \mathbf{V}_e(1) & \mathbf{V}_e(2) & \dots & \mathbf{V}_e(l) \end{bmatrix}. \tag{4.2}$$

Each row i of $\mathbf{M}_e$ illustrates the dissimilarity scores of triplet i across the sequence, and the median of each column j of $\mathbf{M}_e$ is the dissimilarity score of pose I_j and $J_{\psi_{WA\rightarrow WB}(j)}$. Similarly we build an error score matrix $\mathbf{M}'_e$ for alignment $\psi_{WA\rightarrow R}$. $\mathbf{M}_e$ and $\mathbf{M}'_e$ are illustrated visually in Figure 4.2.

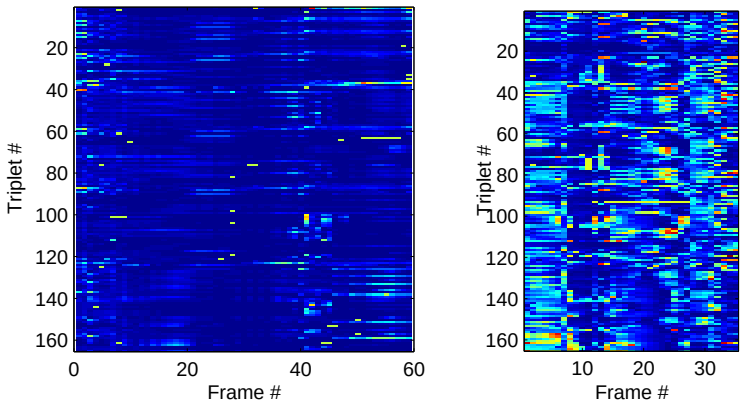
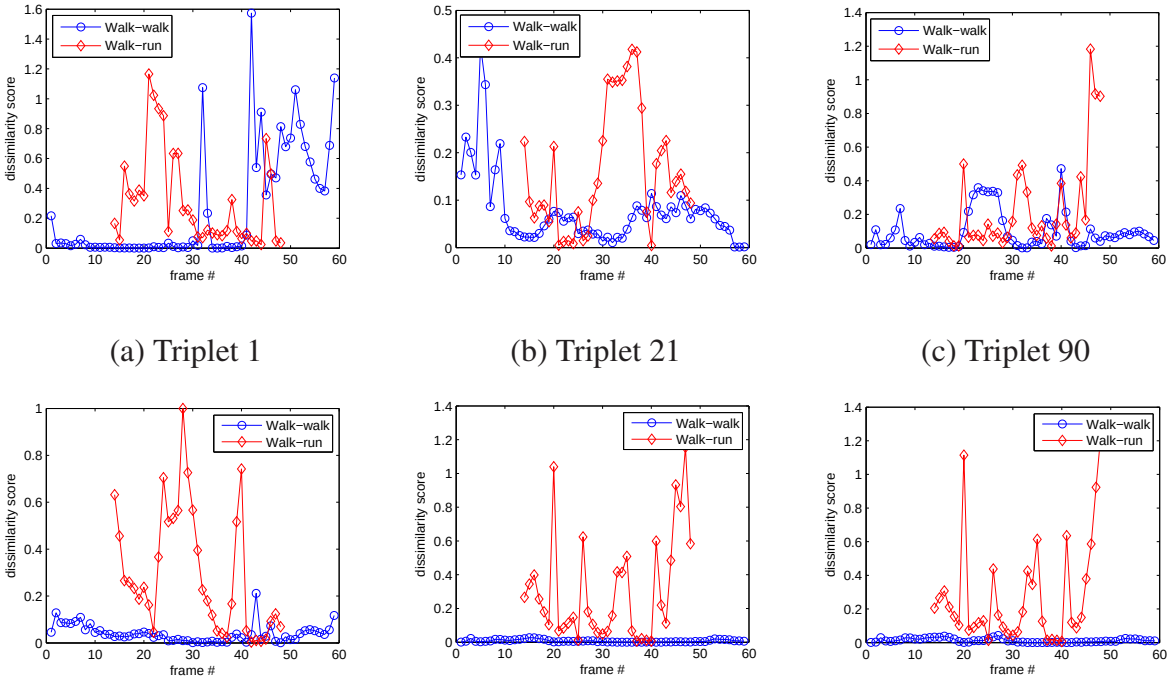


Figure 4.2: Visual illustration of $\mathbf{M}_e$ (left) and $\mathbf{M}'_e$ (right)

To study the role of a triplet i in distinguishing walking and running, we compare the i-th row of $\mathbf{M}_e$ and $\mathbf{M}'_e$, as plotted in Figure 4.3 (a) - (f). We found that, some triplets such as triplets 1 , 21 and 90 have similar error scores in both cases, which means the motion of these triplets are similar in walking and running. On the other hand, triplets 55, 94 and 116 have high error scores in $\mathbf{M}'_e$ and low error scores in $\mathbf{M}_e$, that is, the motion of these triplets in a running sequence is different from their motion in a walking sequence. Triplets 55, 94 and 116 reflect the variation of action walking and running, thus are more informative than triplet 1 , 21 and 90 to the task of distinguishing walking and running action.

In the above experiments, we compare sequences of different actions, and found some triplets that carry more important information of action variations. In the following experiments, we com-



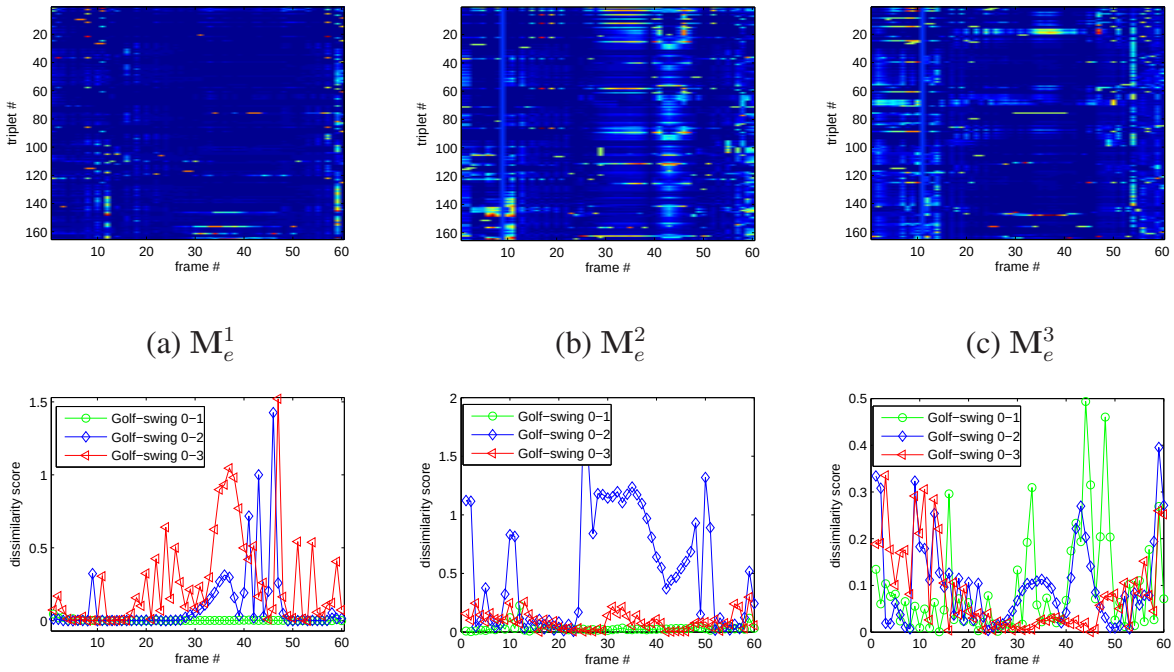
(a) Triplet 1 (b) Triplet 21 (c) Triplet 90

(d) Triplet 55 (e) Triplet 94 (f) Triplet 116

Figure 4.3: Roles of different triplets in action recognition. (a) - (f) are the plots of dissimilarity scores of some triplets across frames in the walk-walk and walk-run alignments.

pare sequences of different individuals performing the same action, and study the roles of triplets in categorizing them in the same group of action: Select four sequences $G0$, $G1$, $G2$, and $G3$ of golf-swing action, and align $G1$, $G2$, and $G3$ to $G0$ using the alignment method described in chapter 3, and then build error score matrix $\mathbf{M}_e^1$, $\mathbf{M}_e^2$, $\mathbf{M}_e^3$ correspondingly as in above experiments. From the illustrations of $\mathbf{M}_e^1$, $\mathbf{M}_e^2$, $\mathbf{M}_e^3$ in Figure 4.4 (a), (b) and (c). The dissimilarity scores of some triplets, such as triplet 120 (see Figure 4.4 (f)) , is very consistent across individuals. Some other triplets such as triplet 20 (Figure 4.4 (d)) and 162 (Figure 4.4 (e)) have various error score

patterns across individuals, that is, such triplets represent the variations of individuals performing the same action.



(a) $\mathbf{M}_e^1$ (b) $\mathbf{M}_e^2$ (c) $\mathbf{M}_e^3$

(d) Triplet 20 (e) Triplet 162 (f) Triplet 120

Figure 4.4: Roles of different triplets in action recognition

Definition 1 *If a triplet reflects the essentials of an action A against other actions, we call it* ***s****ignificant triplet of action A. The other triplets except significant triplets are called* ***t****rivial triplets of action A.*

A typical significant triplet should (1) convey the variations between actions and/or (2) tolerate the individual variations of the same action. For example, triplets 55, 94 and 116 are significant triplets to walking action, and triplet 20 is a significant triplet to golf-swing action.

Intuitively, in the task of action recognition, we should place more focus on the significant triplets while reducing the negative impact of trivial triplets, that is, assigning appropriate influence factor to the body-point triplets. In our approaches of action recognition, this can be achieved by assigning appropriate weights to the similarity errors of body point triplets in equation 2.21. That is, equation 2.21 could be rewritten as:

$$\mathcal{E}(I_1 \rightarrow I_2, J_i \rightarrow J_j) = \underset{\text{all } \Delta_i}{\text{Mean}}(\lambda_i \cdot E(\Delta_i)), \tag{4.3}$$

where $\lambda_1 + \lambda_2 + \ldots + \lambda_T = 1, T = \binom{n}{3}$, n is the number of body points in the human body model.

A question that arises immediately is, how one can determine these weights λ_i of body point triplets in different actions. Manual assignment of weights could be biased and difficult for a large database of actions, and is inefficient when new actions are added in. Automatic assignment of weight values is desired for a robust and efficient action recognition system. To achieve this goal, we propose to use a fixed size dataset of training sequences to learn weight values. Suppose we are given a training dataset $\mathcal{T}$ which consists of $K \times J$ action sequences for J different actions, each of which with K pre-aligned sequences performed by various individuals. Let λ_i^j be the weight value of body joint with label i ($i = 1 \ldots n$) for the action j ($j = 1 \ldots J$). Our goal is to find optimal assignment of λ_i^j which minimize the similarity scores between sequences of different actions and maximize those of same actions. Since the size of the dataset and the alignments of sequences are fixed, this turns out to be an optimization problem on λ_i^j. Our task is to define a good objective function $f(\lambda_1^j, \lambda_2^j, \ldots, \lambda_n)$ for this purpose, and to apply optimization techniques to solve the problem.

4.2.1 Weights on Triplets versus Weights on Body Points

Given a human body model of n points, we could obtain at most $\binom{n}{3}$ triplets, and need to solve a $\binom{n}{3}$ dimensional optimization problem for weights assignment. Even with a simplified human body model of 11 points, this yields a extremely high dimensional ($\binom{11}{3} = 165$ dimensions) problem. On the other hands, the body point triplets are not independent to each other. In fact, adjacent triplets are correlated by their common body points, and the importance of a triplet is also determined by the importance of its three end points (body points). Therefore, instead of using $\binom{n}{3}$ variables for weights of n triplets, we assign n weights $\omega_{1...n}$ to the body points $P_{1...n}$, where:

$$\omega_1 + \omega_2 + \ldots + \omega_n = 1. \tag{4.4}$$

The weight of a triplet $\Delta = \langle P_i, P_j, P_k \rangle$ are then computed as:

$$\lambda_\Delta = \frac{\omega_i + \omega_j + \omega_k}{\binom{n-1}{2}} = \frac{2(\omega_i + \omega_j + \omega_k)}{(n-1)(n-2)}. \tag{4.5}$$

Note that the definition of λ in (4.5) ensures that $\lambda_1 + \lambda_2 + \ldots + \lambda_T = 1$. Based on (4.5), equation (4.3) is rewritten as another form:

$$\mathcal{E}(I_1 \to I_2, J_i \to J_j) = \frac{2}{(n-1)(n-2)} \underset{1 \le i < j < k \le n}{\text{Median}} ((\omega_i + \omega_j + \omega_k) \cdot E(\Delta_{i,j,k})), \tag{4.6}$$

By introducing weights $\{\omega_{1...n}\}$ to body points, we reduce the high dimensional optimization problem to lower dimensional, solvable problem.

4.2.2 Automatic Adjustment of Weights

Before moving on to the automatic adjustment of weights, we first discuss the similarity score of two pre-align sequences. Given two sequences $A = \{I_{1...N}\}$, $B = \{J_{1...M}\}$, and the known alignment $\psi : A \rightarrow B$, the similarity of A and B is:

$$\mathscr{S}(A, B) = \sum_{l=1}^{N} S(l, \psi(l)) = N\tau - N \sum_{l=1}^{N} \mathcal{E}(I_{l \rightarrow r_1}, J_{\psi(l) \rightarrow r_2}), \tag{4.7}$$

where r_1 and r_2 are computed reference poses. To simplify the problem, we replace the median operator in equation (4.6) with mean operator, that is, use the mean triplet error as the dissimilarity score of two pose transitions. Therefore, the proximate similarity score of A and B is:

$$\bar{\mathscr{S}}(A, B) = N\tau - \frac{2 \cdot N}{(n-1)(n-2)} \sum_{l=1}^{N} \sum_{1 \leq i < j < k \leq n} (\omega_i + \omega_j + \omega_k) \cdot E^{l,\psi(l)}(\Delta_{i,j,k}). \tag{4.8}$$

Consider that N, τ, n and $E^{l,\psi(l)}(\Delta_{i,j,k})$ are constants given the alignment ψ, equation (4.8) can be further rewritten into a simpler form:

$$\bar{\mathscr{S}}(A, B) = a_0 - \sum_{i=1}^{n-1} a_i \cdot \omega_i, \tag{4.9}$$

where $\{a_i\}$ are constants computed from (4.8).

Now return to our problem of automatic weights assignment for action recognition. Based on the discussion in the introduction, an good objective function should reflect the intuitions that, significant triplets should be assigned higher weights, while trivial triplets should be assigned lower weights. Suppose we have a training dataset $\mathcal{T}$ which consists of $K \times J$ action sequences for J different actions, each of which with K pre-aligned sequences performed by various individuals.

$\mathcal{T}_k^j$ is the k-th sequence in the group of action j, and $\mathcal{R}^j$ is the reference sequence of action j. To find the optimal weights assignment for action j, we define the objective function as:

$$f^j(\omega_1, \omega_2, \ldots, \omega_{n-1}) = \mathcal{Q}_1 - \alpha\mathcal{Q}_2 - \beta\mathcal{Q}_3, \tag{4.10}$$

where α and β are non-negative constants and

$$\mathcal{Q}_1 = \frac{1}{K}\sum_{k=1}^{K}\bar{\mathscr{S}}(\mathcal{R}^j, \mathcal{T}_k^j), \tag{4.11}$$

$$\mathcal{Q}_2 = \frac{1}{K}\sum_{k=1}^{K}\bar{\mathscr{S}}(\mathcal{R}^j, \mathcal{T}_k^j)^2 - \frac{1}{K^2}\left(\sum_{k=1}^{K}\bar{\mathscr{S}}(\mathcal{R}^j, \mathcal{T}_k^j)\right)^2, \tag{4.12}$$

$$\mathcal{Q}_3 = \frac{1}{K(J-1)}\sum_{1\le i\le J, i\neq j}\sum_{k=1}^{K}\bar{\mathscr{S}}(\mathcal{R}^j, \mathcal{T}_k^i). \tag{4.13}$$

The optimal weights for action j is then computed by minimizing $f^j(\omega_1, \omega_2, \ldots, \omega_{n-1})$:

$$\langle\omega_1, \omega_2, \ldots, \omega_{n-1}\rangle = \underset{\omega_1,\omega_2,\ldots,\omega_{n-1}}{\operatorname{argmax}} f^j(\omega_1, \omega_2, \ldots, \omega_{n-1}). \tag{4.14}$$

In the objective function, we use $\mathcal{T}_1^j$ as the reference sequence for action j, and the term $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are the mean and variance of similarity scores between $\mathcal{T}_1^j$ and other sequences in the same action. $\mathcal{Q}_3$ is the mean of similarity scores between $\mathcal{T}_1^j$ and all sequences in other different actions. The intuition of $f^j(\omega_1, \omega_2, \ldots, \omega_{n-1})$ is to achieve high similarity scores for all sequences of same action j, and low similarity scores for sequences of different actions. The second term $\mathcal{Q}_2$ is to ensure the consistency of sequences in the same group.

As $\mathcal{Q}_1$ and $\mathcal{Q}_3$ are linear functions, and $\mathcal{Q}_2$ is quadratic polynomial, our objective function $f^j(\omega_1, \omega_2, \ldots, \omega_{n-1})$ is quadratic polynomial function, and the optimization problem is actually

a quadratic programming (QP) problem. There are a variety of methods solving QP problem, including Interior point, active set, conjugate gradient, etc. The properties and solutions of QP problem have been well studied in numerical optimization research. In our problem, we adapt the conjugate gradient method, with the initial value at $\langle \frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n} \rangle$.

4.3 Experiments

In this section, we apply the proposed weighting based approach to the action recognition problem, and compare its performance with non-weighting methods proposed in chapter 3. For comparison, we use the same MoCap testing data as in chapter 3, and build a MoCap training dataset which consists of totally $2 \times 17 \times 5 = 170$ sequences for 5 actions (walk, jump, golf swing, run, and climb): each action is performed by 2 actors, and each instance of action is observed by 17 cameras set up as in Fig. 3.3. The same 11-point human body model in Figure 2.1 is adapted in the training data. We use the same set of reference sequences for the 5 actions, and align the sequences in the training set against the reference sequences.

To obtain optimal weighting for each action j, we first align all testing sequences against the reference sequence $\mathcal{R}^j$, and save of the similarity scores of triplets for each pair of matched poses. The objective function $f^j(\omega_1, \omega_2, \ldots, \omega_{10})$ is then built based on equation (4.10), and the computed similarity scores of triplets in the alignments. $f^j(\cdot)$ is a 10-variate function, and the weights ω_i are

constrained by

$$\begin{cases} 0 \leq \omega_i \leq 1, i = 1 \ldots 10, \\ \sum_{i=1}^{10} \omega_i \leq 1. \end{cases} \tag{4.15}$$

The optimal weights $\langle \omega_1, \omega_2, \ldots, \omega_1 0 \rangle$ are then searched to maximize $f^j(\cdot)$, with the initial location at $\langle \frac{1}{11}, \frac{1}{11}, \ldots, \frac{1}{11} \rangle$. The conjugate gradient method is then applied to solve this optimization problem.

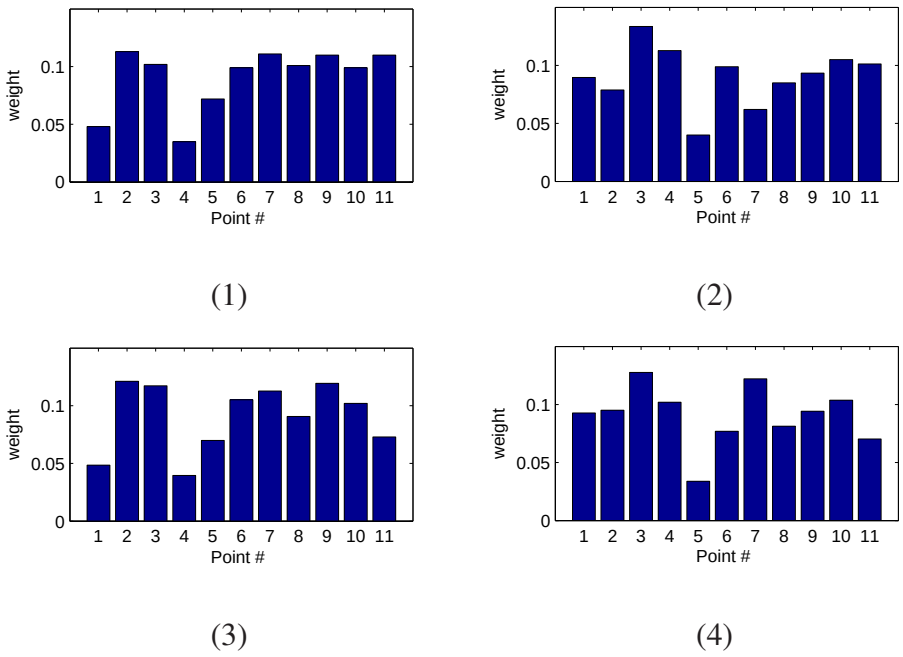


Figure 4.5: Examples of computed weights. (1) and (2) are computed weights for walking and jumping correspondingly based on fundamental ratio invariant; (3) and (4) are computed weights for walking and jumping correspondingly based on eigenvalue equality invariant.

After performing the above process for all the actions, we obtain a set of weights $\mathcal{W}^j$ for each action j. We carry out the experiments with both fundamental ratio invariant method and eigenvalue equality method. As these methods behave differently, the objective functions we obtain

Table 4.1: Confusion matrix using eigenvalue equality invariant: Large values on the diagonal entries indicate accuracy.

Ground-truth	Recognized as				
	Walk	Jump	Golf Swing	Run	Climb
Walk	46	1		1	2
Jump	1	48			1
Golf Swing	1		48	1	
Run		2		48	
Climb	4	1	1		44

contain different sets of coefficients. Figure 4.5 shows the computed weights for walking and jumping when using different geometric invariants. The weights we obtained based on fundamental ratio invariant method is different from those we obtained based on eigenvalue equality method. However, as shown in the figure, some triplets have relatively high weights in both results.

The signature of each action in our action recognition system now is composed of two parts: a reference sequence $\mathcal{R}^j$ and a weighting $\mathcal{W}^j$. We denote this weighted signature as $\mathcal{WR}^j = \{\mathcal{R}^j, \mathcal{W}^j\}$. The action recognition system is now based on the weighted similarity score in equation (eqn:weightedSimilarity). We perform the same action recognition experiments based on MoCap data as described in section 3.2.4.1, with the results reported in Table 4.1 and 4.2. The overall recognition rate is 93.6% using eigenvalue equality invariant, and 90.4% using fundamental ratio

Table 4.2: Confusion matrix using fundamental ratios invariant: Large values on the diagonal entries indicate accuracy.

Ground-truth	Recognized as				
	Walk	Jump	Golf Swing	Run	Climb
Walk	45	1	1	2	1
Jump	2	47		1	
Golf Swing		1	47	1	1
Run	3	1		45	1
Climb	4	1	1	2	42

invariant, which are improved by 2% and 8.8% correspondingly compared to the un-weighted system.

5 HUMAN ACTION STYLE ANALYSIS

Human action style analysis is another important area in interpreting human motion, which is motivated by the need of various applications such as surveillance systems, ergonomic evaluations, etc. Its goal is to study the style variations of the individuals in performing the same action, and to discover internal factors such as age, gender, and human identity, that result in these variations.

The problem of action style analysis is related to the action recognition problem in that, the target of an action recognition system is to find the features that distinguish different actions, while in action style analysis, stylistic features, e.g., stride parameters of walking gaits, are extracted from instances of the same action to reflect the style variations of individuals. Therefore, theoretic results in action recognition can be borrowed to the study of action style analysis.

5.1 Related works

It has been proven that humans can recognize actions from limited types of input such as point lights and low quality video [BD01b, JOH73]. Even with such limited information, we are capable to differentiate stylistic action differences, such as the gender [BCK78, PLR02] and age [Dav01] of the walker. There has been recent work on the study of action style variation in com-

puter vision. Wilson and Bobick [WB99] use a Parameterized-HMM to model spatial pointing gestures by adding a global variation parameters in the output probabilities of the HMM states. In [TW] Tenenbaum et al. use a bilinear model to separate perceptual content and style parameters. Davis [Dav01] proposed an approach to determine age of people based on variations in relative stride length and stride frequency over various walking speeds. In [DT02] Davis and Taylor use regularities in walking to classify typical from atypical gaits. Davis et al. [DG04] presented a three-mode (body pose, time, and style) expressive-feature model for representing and recognizing performance styles of human actions. The application of style analysis in computer animation for generating new animation styles of human motion have also been reported in [UAT95, BH00, Vas02, VT02, DK02, CCZ00].

Gait recognition is another problem related to action style analysis, which has very important implications for different domains such as surveillance, medical diagnosis, etc. It is based on the widely held belief that humans can distinguish between gait patterns of different individuals, by examining gait properties such as stride length, bounce, rhythm, and speed, etc. An early report on the ability to recognize people from gait was presented by Beardsworth and Buckner [BB81]. They showed that the ability to recognize oneself from point-light features is greater than that to recognize others. The ability of people to identify others using gait information alone has also been supported by the studies of Stevenage et al. [SNV99] and Schollhorn et al. [SNS02]. Existing approaches in gait recognition can be widely grouped into two categories: model based approaches [LG02, Tro02, MDK64] and model free approaches [BCD02, HHN99, LB98, KCY03, SRC03, Rab89, TC03].

Like in other problems of human motion analysis, view invariance is an important requirement of gait recognition system. However, only a few work in the literature take into account the view invariance. Shakhnarovich et al. [SLD01] propose an approach that integrates face and gait recognition from multiple views. As other work that use multiple view data, their approach is limited to the number of views being used and is not "true" view-invariant. Kale et al.[KCC03] propose a view-invariant method for the case when the person is far from the camera. They synthesize a side view from any other arbitrary view using a single camera, and apply methods based on side view of walking to solve the gait recognition problem.

5.2 Human Action Style Analysis Using Body Point Triplets

As proven by the success of using body point triplets in human action recognition in chapter 3, the body point triplets model is an effective representation of non-rigid human motion. Both the similarity and irregularity of human motions are reflected by the motion of different human body points, however, studying body points individually under multiple views scenarios is usually difficult due to the lack of information in a single body joint. By decomposing a body pose into body point triplets, or in another word, composing every three body points into a triplet, we obtain elements that provide sufficient information for geometry analysis, while conveying the essentials of human motion at atomic level. From our study of human pose and action recognition, we have observed that body point triplets carry various information about human motion, for instance, some triplets may convey the differences of actions, while some other triplets may reflect the variations

of individuals in performing the same actions. The roles of these triplets are not independent, and they are usually studied as a group to understand the human motion.

In chapter 3, we use all triplets to calculate similarity of human actions, in which the variations of different actions are studied. When the studied sequences are in the same category of human action, variations of individual can be recognized by examining the motion of specific body point triplets. For example, in the case of studying two sequences of male and female walking, the motion of the body point triplet that consists of hip, knee and foot usually appears differently between two sequences, due to the different styles of swagger and body swing between male and female. Therefore, it is feasible to derive style features of individuals from the motion of such body point triplets, and apply them to various tasks of understanding human motions, e.g., gender identification from human motion.

In general, a good style feature should have the following properties:

1. It should be able to convey the variations of styles/individuals.

2. It should be consistent under different camera parameters and viewpoints, that is, it should be view invariant.

In the Section 5.2.1, we will introduce our representation of style information based on body point triplets, and discuss its properties in the above aspects.

5.2.1 Triplet-based Representation of Human Action Styles

A major task of action style analysis is to represent the style information from the motion sequences, and then extract style features for classification. We can consider using the motion patterns of triplets, since the style variations of human motion are reflected by the motion of body point triplets. Consider the case of walking action, if we align a walking sequence of some individual against the reference sequence, the motion patterns of triplets are reflected in the similarity scores between two sequences through the whole sequences. This provides us a direction to derive stylistic features from the motion sequences.

Suppose we are given a reference sequence $\mathcal{R}$ of length n, and a target sequence $\mathcal{T}$ of length m, in the group of action A. We can find an alignment or mapping, $\{\psi : \mathcal{T} \rightarrow \mathcal{R}\}$, based on the sequence alignment algorithm described in section 3.2.3. As we discuss in section 3.2.3, the selection of constant threshold τ will affect the alignment results. A higher valued τ has more tolerance of pose differences, and therefore results in a longer alignment path. Thus, by setting an appropriate global threshold τ for all sequences in action A, we can always find a full mapping from $\mathcal{T}$ to $\mathcal{R}$. Since τ is set globally for all sequences in action A, it does not change the correlation of similarity scores of different sequences. In the analysis of human action styles, we are more concerned of the correlation of score values between sequences, in stead of score values themselves.

From the alignment $\{\psi : \mathcal{T} \rightarrow \mathcal{R}\}$, we can build two matrices that reflect the action style in the target sequence $\mathcal{T}$:

Triplet Variation Matrix (TVM) As described in 4.2, for alignment $\{\psi : \mathcal{T} \rightarrow \mathcal{R}\}$ we can build an error score matrix $\mathbf{M}_e$. Each row of $\mathbf{M}_e$ illustrates the dissimilarity scores of corresponding triplet across the sequence of $\mathcal{T}$ and each column is the similarity score vector of pose I in $\mathcal{R}$ and its matched pose $\psi(I)$ in $\mathcal{T}$. $\mathbf{M}_e$ reflects the variations of all triplets in $\mathcal{T}$ compared with the reference $\mathcal{R}$. A peak or valley in $\mathbf{M}_e$ means that the motions of corresponding triplets at the specific poses appear very different in $\mathcal{T}$ from those in $\mathcal{R}$. To reduce the noise introduced by degenerated or close-to-degenerated triplets, we apply a median filter to $\mathbf{M}_e$:

$$\hat{\mathbf{M}}_e = \mathrm{MedianFilter}(\mathbf{M}_e), \tag{5.1}$$

where $\mathrm{MedianFilter}(\cdot)$ is a $M \times M$ median filtering operator. We then resize $\hat{\mathbf{M}}_e$ to a $n \times n$ matrix $\mathbf{M}_{tv}$, which is named as *Triplet Variation Matrix*, by applying bilinear interpolation on $\hat{\mathbf{M}}_e$.

Poses Variation Matrix (PVM) During aligning sequence $\mathcal{R}$ and $\mathcal{T}$, we apply dynamic programming on the pose similarity matrix $\mathcal{S}$, in which $\mathcal{S}[i, j]$ is the similarity scores of pose i of $\mathcal{R}$ and pose j of $\mathcal{T}$. In fact, $\mathcal{S}[i, j]$ describes the pose-to-pose correlation between the two sequences. When $\mathcal{T}$ and $\mathcal{R}$ are the same sequence, $\mathcal{S}$ is a special case of the self-similarity matrix, which is used for action recognition [JDL08]. The patterns of element values in $\mathcal{S}[i, j]$ representations of the target sequence's properties, and can be used to describe the style variation of the target sequence. Based on $\mathcal{S}$, we define the *pose variation matrix* as:

$$\mathbf{M}_{pv} = \mathrm{BilinearInterp}(\mathrm{GaussianFilter}(S)), \tag{5.2}$$

where $\mathrm{GaussianFilter}(\cdot)$ is $N \times N$ Gaussian filtering operator, and $\mathrm{BilinearInterp}(\mathbf{M})$ is resizing $\mathbf{M}$ to a $n \times n$ matrix by bilinear interpolation.

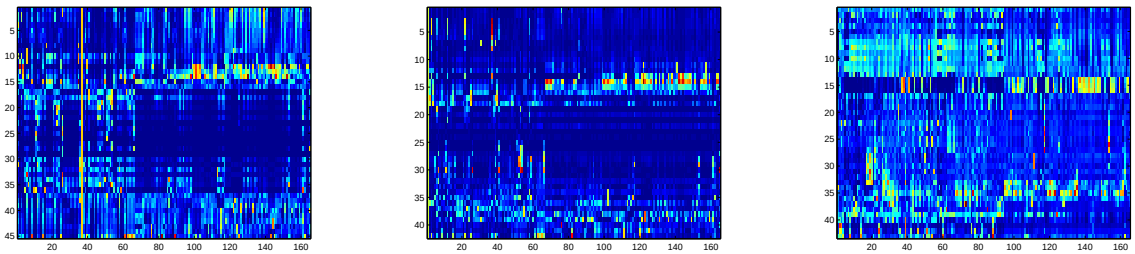
TVM and PVM describe the style variations of a sequence in different aspects: TVM illustrates the lower level (body point triplet level) of style variations, while PVM represents the motion style from a higher view point by looking at the whole body pose.

In the following sections, we will discuss the properties of TVM and PVM. Without loss of generality, we study the action of kicking as an example in the following sections. We assume that sequences of kicking action are already segmented from video data, that is, the start and end frames are provided for all kicking sequences. In our study, we use the IXMAS dataset [WRB06] in which videos of 13 actions are captured under 5 cameras, and each action is performed by 11 actors for 3 times/instances. For kicking action, we use "bao1" sequence from camera 2 as reference sequence.

Figure 5.1: Example sequences in IXMAS dataset.

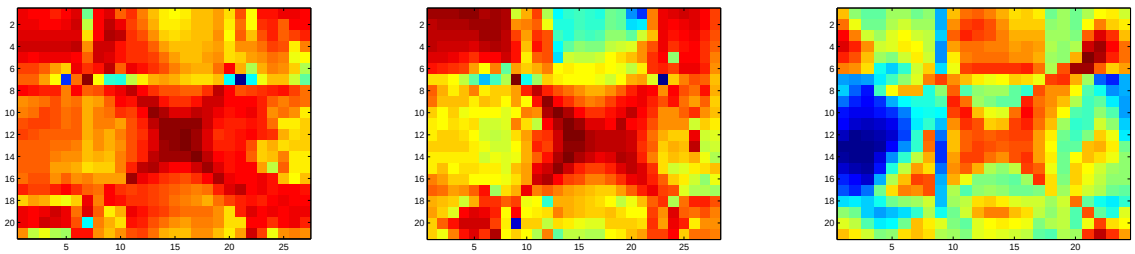
5.2.1.1 Variations of Individuals

We select 3 kicking sequences performed by 3 actors in the data set, align them to the reference sequence and compute the corresponding TVM and PVM for each sequence (see Figure 5.2 and 5.3).



(a) (b) (c)

Figure 5.2: TVMs of sequences performed by different actors: (a) is associated to actor ”Florian”, (b) to ”Nicolas”, and (c) to ”Srikumar”



(a) (b) (c)

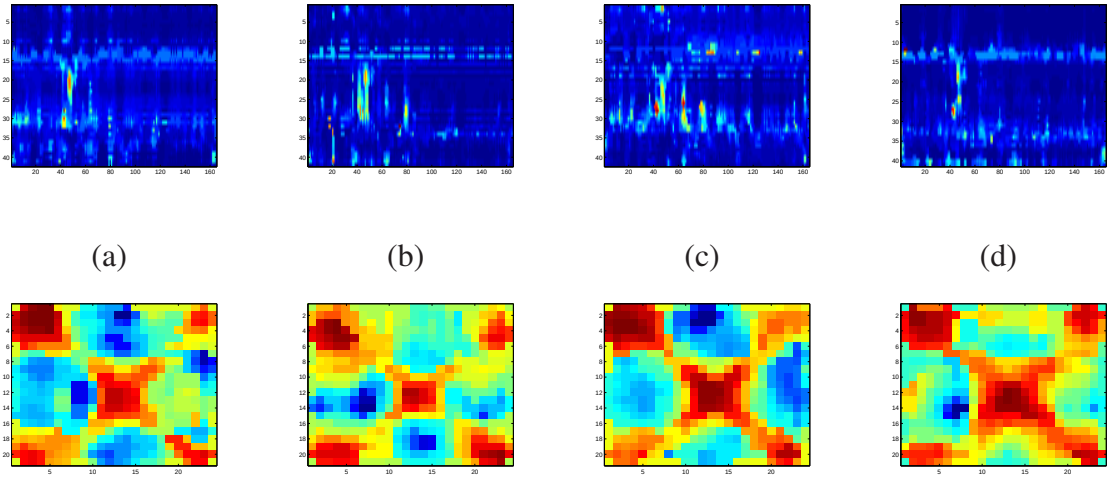
Figure 5.3: PVMs of sequences performed by different actors: (a) is associated to actor ”Florian”, (b) to ”Nicolas”, and (c) to ”Srikumar”

As shown in Figure 5.2 and 5.3, the TVM and PVM have different patterns in the three sequences. The different locations of peaks and valleys in the TVM plots suggest that the corre-

sponding triplets move differently at various time spots in these sequences. The PVM also provide good illustration of different styles of these sequences, though their patterns in the diagonal look similar due to the same action. The above illustrations set a good example that TVM and PVM effectively represent the different styles of individuals.

5.2.1.2 View Invariance

To study the consistency of TVM and PVM in different viewpoint and camera parameters, we arbitrarily select one instance of an actor and its captured videos by 4 different cameras. Similarly we compute the TVM and PVM for each sequence, and plot them in Figure 5.4.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5.4: TVM and PVM under different viewpoints. (a) - (d) are illustrates of TVMs and (e) - h) are PVMs that correspond to camera 1 - 4.

As shown in Figure 5.4, the computed TVM for each actor under various camera setups have similar peaks and valleys. Though minor variations exist, it is still easy to distinguish individuals from the visual difference of TVMs. The same observation is made regarding the PVMs. These observations show that as expected TVM and PVM are view invariant, since they are derived based on geometric invariants.

From the above analysis, we found the our triplet variance matrix and pose variance matrix are effective in representing the stylistic features of human motion, and are invariant to camera intrinsic parameters and viewpoints.

5.2.2 Gender Recognition Based On Action Stylistic Information

In this section, we discuss the application of gender recognition from action style representation TVM and PVM. We are provided with a set of training sequences which are labelled as w_0 (female) and w_1 (male). Our goal is to find a classifier that correctly categorizes any input action sequence T as w_0 or w_1.

As discussed in previous sections, our TVM and PVM provides a good representation of action style of the sequence, therefore can be utilized to represent the sequences for our task. We can serialize TVM and PVM to a vector to represent a sample, however, such vector is high dimensional, which makes the problem quite challenging. To make the task more suitable to solve, we first reduce the dimensionality of data by downsampling $\mathbf{TVM}$ to a $d_1 \times d_1$ matrix $\mathbf{TVM}'$ and $\mathbf{PVM}$ to

a $d_2 \times d_2$ matrix $\mathbf{PVM}'$. $\mathbf{TVM}'$ and $\mathbf{PVM}'$ are then serialized as two vectors, which are stacked into a single $d = d_1^2 + d_2^2$ dimensional vector

$$\mathbf{x} = \left[\begin{array}{cccccccc} \mathbf{PVM}'_{11} & \mathbf{PVM}'_{12} & \ldots & \mathbf{PVM}'_{d_1 d_1} & \mathbf{TVM}'_{11} & \mathbf{TVM}'_{12} & \ldots & \mathbf{TVM}'_{d_2 d_2} \end{array} \right]^T . \quad (5.3)$$

During the downsampling, we need to select sufficient large d_1 and d_2 so that necessary details in TVM and PVM are retained. Therefore, dimension reduction through downsampling is quite limited. Moreover, due to irregularity of human motion, the same action style may be performed slightly different in various instances, thus produce different patterns of triplet motion or pose variation on the TVM and PVM. Some of these patterns may be essential to the human action styles, while the others are merely random and trivial 'noises' to the task of action style analysis. Therefore, for a better interpret of the style of human motions, we need to extract more robust and reliable stylistic features from the TVM and PVM.

5.2.2.1 Gender Classification based on Principal Component Analysis

In order to reveal the underlying stylistic information in the TVM and PVM, we need to describe the data in a way that critical and trivial triplet or pose patterns are better separated. Principal Component Analysis (PCA) provides a good solution for this purpose. PCA is one of the most popular technique for feature selection and dimensionality reduction. It provides an optimal data representation in the mean square error sense, by seeking principal components in the feature

space. PCA has been proven successful in diverse applications from neuroscience to computer visions, such as animation [BH00, GBT04], face recognition [TP91], etc.

Suppose we have a set of N d-dimensional samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$. The goal of PCA is to find a nature set of d' orthonormal basis vectors $\{\mathbf{e}_i\}$ to represent the samples such that the criterion function

$$\mathcal{J}_{d'} = \sum_{k=1}^{N} \left\| \left(\mathbf{m} + \sum_{i=1}^{d'} a_i^k \mathbf{e}_i - \mathbf{x}_k \right) \right\|^2 \tag{5.4}$$

is minimized, where $\mathbf{m}$ is the sample mean,

$$\mathbf{m} = \frac{1}{n} \sum_{k=1}^{N} \mathbf{x}_k, \tag{5.5}$$

and

$$\mathbf{a}^k = [\ a_1^k \quad a_2^k \quad \ldots \quad a_{d'}^k\ ]^T \tag{5.6}$$

are defined as principal components. The solution is to compute the eigenvalues and eigenvectors of the scatter matrix $\mathbf{S}$

$$\mathbf{S} = \sum_{k=1}^{N} (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T, \tag{5.7}$$

and sort the eigenvalues and eigenvectors according to decreasing eigenvalue. The d' largest eigenvectors are then use as the basis vectors $\{\mathbf{e}_i\}$, that is, $\mathbf{e}_i$ corresponds to the i-th largest eigenvalue λ_i. Usually d' is much lower than d, which implies that the d' dimensions are inherent subspaces that govern that samples, while the remaining $d - d'$ dimensions are merely noises. A sample $\mathbf{x}_k$ can now be represented by principal components though projecting onto the d' dimensional subspace as $\tilde{\mathbf{x}}_k$:

$$\tilde{\mathbf{x}}_k = \mathbf{A}^T(\mathbf{x}_k - \mathbf{m}), \tag{5.8}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \dots & \mathbf{e}_{d'} \end{bmatrix}. \tag{5.9}$$

The basis vectors computed by PCA are in the direction of the largest variance of the training vectors, and they convey the stylistic elements inherent to the specific motion. We call these basis as eigenstyles. These eigenstyles span a style space of the specific motion. When a sample $\mathbf{x}_k$ is projected onto the style space, its vector $\tilde{\mathbf{x}}_k$ describes the significance of these eigenstyles in the sample. We therefore define $\tilde{\mathbf{x}}_k$ as the stylistic feature of the sample. A style sample/representation can be reconstructed with some error based on the eigenstyles and its stylistic feature from equation. (5.8).

We adapt the k-nearest neighbor algorithm to classify sequences represented by the PCA based stylistic feature. Suppose we are provided with a set of stylistic feature vectors $\{\tilde{\mathbf{x}}_k | k = 1, 2, \dots, n\}$ after PCA, and their corresponding labels $\{\mathcal{L}_k | k = 1, 2, \dots, n, \mathbf{L}_k \in \{w_0, w_1\}\}$. A target sequence T is classified as w_0 or w_1 based on the following procedure:

1. The PVM and TVM of sequence T are first computed, and then downsampled to a d-dimensional vector $\mathbf{x}$.

2. $\mathbf{x}$ is projected onto the eigenstyle space as $\tilde{\mathbf{x}}$.

3. The Euclidian distances between $\tilde{\mathbf{x}}$ and all $\{\tilde{\mathbf{x}}_k$ in the training set are computed, and T is classified as w_i which is most frequent among the k training vectors nearest to $\tilde{\mathbf{x}}$, where k is the closest odd integer to $\sqrt{n}$.

5.2.2.2 Gender Classification using Fisher Discriminant Analysis

The PCA method find eigenstyles to describe as much variance of the data as possible, and provide good features to describe the data. However, these eigenstyles are not necessary to be useful for discriminating between data in different classes. The $d - d'$ dimensions that are thrown away in PCA may still contain useful information for our classification task. For our task, we need to extract features for classification, which is different from extracting features for describing data.

PCA is an unsupervised technique that seeks features which are efficient for describing data, and it does not utilize the label information of data. Unlike PCA, discriminant analysis seeks features that are efficient to discriminate the classes given the labelled data. To this purpose, suppose the data $\{\mathbf{x}_i | i = 1 \dots n\}$ are categorized into w_0 and w_1, Fisher Linear Discriminant Analysis project the data $\mathbf{x}_i$ to point y on a line $\mathbf{w}$ by a linear combination of the components of $\mathbf{x}$:

$$y = \mathbf{w}^T \mathbf{x}, \tag{5.10}$$

and seek an optimal $\mathbf{w}$ that results in best separation between points with different labels. It is solved by maximizing the objective function:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}, \tag{5.11}$$

where the within-class scatter matrix $\mathbf{S}_W$ is defined as

$$\mathbf{S}_W = \sum_{\mathbf{x} \in w_0} (\mathbf{x} - \mathbf{m}_0)(\mathbf{x} - \mathbf{m}_0)^T + \sum_{\mathbf{x} \in w_1} (\mathbf{x} - \mathbf{m}_1)(\mathbf{x} - \mathbf{m}_1)^T, \tag{5.12}$$

the between-class scatter matrix $\mathbf{S}_B$ is defined as

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_0)(\mathbf{m}_1 - \mathbf{m}_0)^T, \tag{5.13}$$

and

$$\mathbf{m}_i = \frac{1}{n} \sum_{\mathbf{x}_i \in w_i} \mathbf{x}_i. \tag{5.14}$$

As discussed in [DHS01], $J(\cdot)$ is independent of $\|\mathbf{w}\|$, and the solution of $\mathbf{w}$ that optimized $J(\cdot)$ is

$$\mathbf{w} = \mathbf{S}_w^{-1}(\mathbf{m}_0 - \mathbf{m}_1). \tag{5.15}$$

Through $\mathbf{w}$ we project our style vectors $\mathbf{x}$ on a line, and the projected scalar value y is our extracted stylistic feature. We thus convert the d-dimensional classification problem to a hopefully more manageable one-dimensional one. All that remains for our task to find a threshold that separate the projected points into w_0 and w_1. Here the decision surface is reduced to a scalar value. We assume that the stylistic vectors of both classes exhibits approximately the same distributions, therefore we choose the separate threshold as

$$c = \mathbf{w}(\frac{\mathbf{m}_1 + \mathbf{m}_2}{2}). \tag{5.16}$$

For our problem, linear discriminant is sufficient as shown from experiments in the next section. However, when linear discriminant is not complex enough, FDA can be extended to nonlinear discriminant by the Kernel FDA[].

5.3 Experiments

In this section, we present experiments to demonstrate the effectiveness of our proposed gender recognition from human motion based on the stylistic features. We test our methods on two actions, kicking and walking for gender recognition. Our experiments are based on IXMAS dataset,

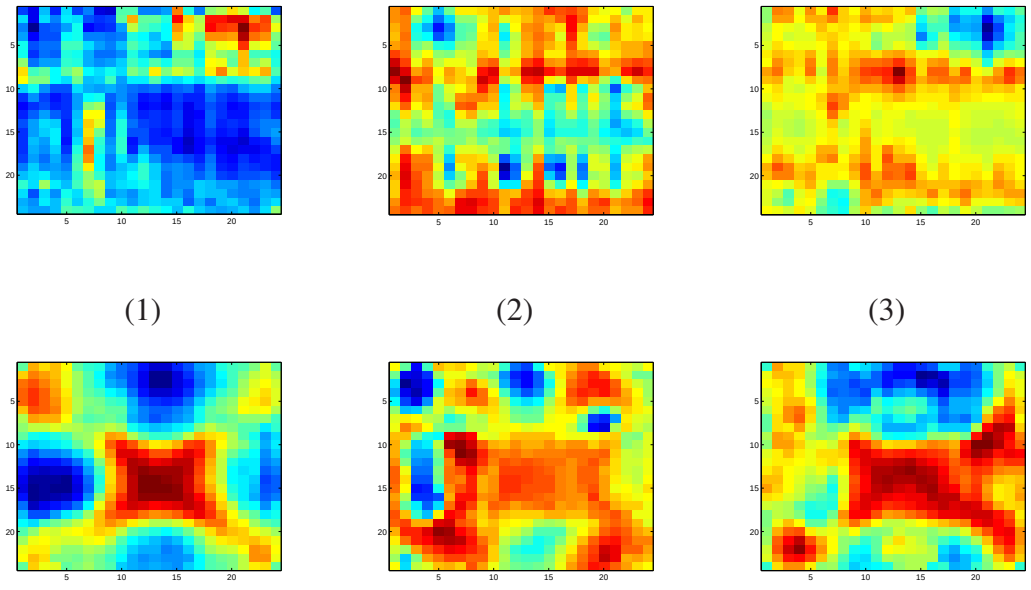
Table 5.1: Male and female classification rates with different d' for kicking action

d'	1	2	3	4	5	6	7	8	9	10	11	12	13
Male classification rate	.540	.619	.635	.730	.778	.778	.841	.857	.873	.873	.873	.873	.873
Female classification rate	.492	.571	.603	.683	.762	.810	.825	.825	.873	.873	.873	.889	.889

which consists of 13 daily-live motions performed each 3 times by 11 actors at freely position and orientation, and observed from 5 cameras setup at various viewpoints. For all sequences of the selected actions, we assume that body points have been tracked, and the actions are already segmented. We arbitrarily select one of the actors, e.g. "Pao", as reference for each action, and the rest actors for training and testing. All the sequences are downsampled to 30 fps for simplicity.

For each action, we select a number of sequences as a training set, which include 2 female actors and 2 male actors, and each actor perform the action 2 times while observed by camera position 1, 2 and 3. We therefore have 24 sequences in our training set for each action, and the rest 126 sequences are used as testing sequences . For each action, the PCA based method and FLDA method are applied to classify the testing sequences as male or female. We downsample both computed PVM and TVM to 24×24 matrices, thus the input vectors for PCA and FLDA are 1151 dimensional.

Figure 5.5 displays the first 3 eigenstyles we obtained for kicking action. We use different values d' when using our PCA method, and measure their resulted classification rates, as presented in table 5.1 and 5.2, and also in figure 5.6.



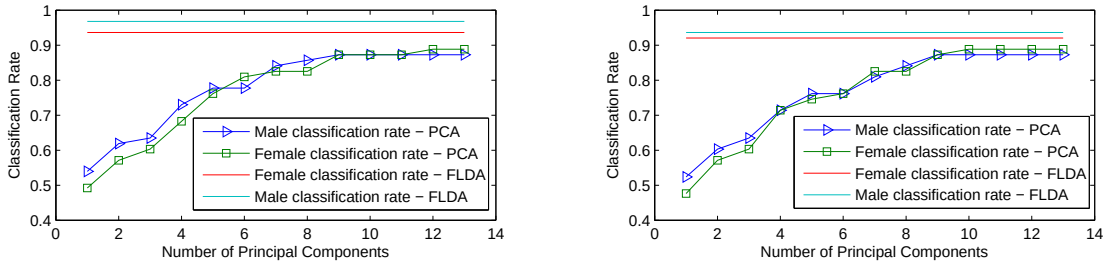
(1) (2) (3)

(4) (5) (6)

Figure 5.5: The first 3 eigenstyles computed for kicking action. (1) - (3) are the TVM parts of the eigenstyles, while (4) - (6) are the PVM parts.

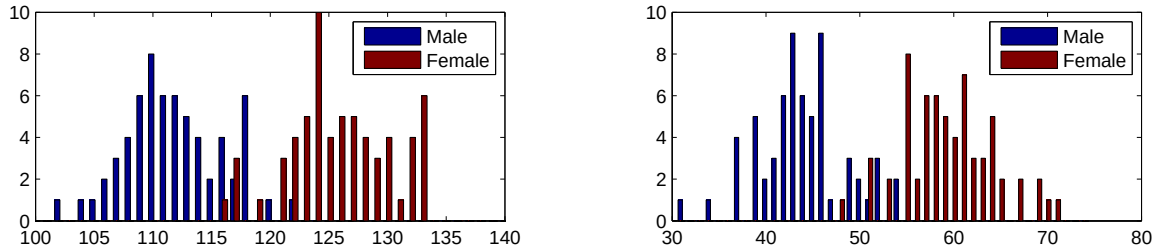
Table 5.2: Male and female classification rates with different d' for walking action

d'	1	2	3	4	5	6	7	8	9	10	11
Male classification rate	.524	.603	.635	.714	.762	.810	.841	.873	.873	.873	.873
Female classification rate	.476	.571	.603	.714	.746	.762	.825	.825	.873	.889	.889



(a) (b)

Figure 5.6: Classification rates using PCA method and FLDA method. (a) and (b) illustrate the values classification rate with different d' for kicking and walking action respectively.



(a) (b)

Figure 5.7: Distribution of projected points of stylistic vectors. (a) is for kicking 118.36 and (b) for walking action 72.175.

With FLDA method, we project the high dimensional data into one dimension. We illustrate the distribution of the projected points by histograms in Figure 5.7. As illustrated from the histograms, the data of two classes are well separated for kicking, and for walking there are more overlapped regions, but the separation is still good for classification task. For kicking action, the computed threshold is 118.360, and the female and male classification rate is 93.65% and 96.83% respectively. The threshold for walking action is 72.175, and the female and male classification is 93.65% and 92.06%.

We plot the resulted classification rates that based on FLDA method with those based on PCA method in Figure 5.6. As suggested by the results, the FLDA method is more efficient in the task of gender recognition, partially due to the fact that FLDA method better utilizes the labels of the training data, and exact features that are more efficient to the discriminant task.

6 CONCLUSIONS

In this dissertation we study the geometric invariance in human motion and its various applications in understanding human motion. We first present a novel technique to analyze the human motion. By representing the human body as a set of points, we decompose a body posture into a set of triplets (planes). The non-rigid motion of human body is therefore transformed to the rigid motion of world planes defined by body point triplets. Thus the analysis of non-rigid human motion reduces to the study of triplet motions, the geometric properties of which have been well studied in multiple view geometry. Using this technique, we study the problem of measuring similarity of human pose transitions, and proposed two different types of geometric invariants in the human motion. The first geometric invariant we propose is the *Fundamental Ratios*. A moving plane observed by a fixed camera induces a fundamental matrix $\mathbf{F}$ across multiple frames, where the ratios among the elements in the upper left 2×2 sub-matrix are herein referred to as the *Fundamental Ratios*. We show that fundamental ratios are invariant to camera internal parameters and orientation, and hence can be used to identify similar plane motions, and further similar human motions from varying viewpoints. The second geometric invariance we propose is the eigenvalues equality, which is based on the fact that the homography induced by the motion of a triplet of body points in two identical *pose transitions* reduces to the special case of a homology. We exploit the equality of

two of its eigenvalues to impose constraints on the similarity of the *pose transitions* between two subjects, observed by different perspective cameras and from different viewpoints.

We then demonstrate the application of our proposed geometric invariance to pose detection and action recognition, which are two major problems of human motion analysis. We propose a new template-based approach for view-invariant recognition of body poses in video sequences. Our templates for specific body poses are extracted from video data, and in addition to spatial information encode temporal information of *body pose transitions*. The two geometric invariants proposed earlier provide us with two different solutions for view-invariant detection of human poses from video data. Extensive experimental results show that our method can accurately identify human poses from video sequences when they are observed from totally different viewpoints with different camera parameters. Next we extend the pose recognition problem to aligning two video sequences of the same human action, based on which we proposed a novel technique of view-invariant recognition of human actions. Instead of regarding an action sequence as a sequence of body poses like existing approaches, we regard it as a sequence of pose transitions, and utilize our proposed geometric invariance to align and match two action sequence. Experiments results evaluated over semi-synthetic data obtained from motion capture database and real video data confirm that our methods can recognize human postures and actions under substantial amount of noise, even when the viewpoints and camera parameters are unknown and totally different. We study the roles of body point triplets in the understanding human motions, and proposed a weighting based extension of our action recognition approaches.

We also study the triplet motion variations and pose variations in the same class of action, and utilize them as representations of the stylistic information of human actions. Through PCA and FLDA we extract stylistic features from these representations, and apply them in the gender recognition problem. As validated by the experimental results, our stylistic features reflect the inherent differences between actions performed by male and female.

As proven by our research, the triplet based analysis has a very high potential in solving various problems in understanding human motions. The variations of triplet motion convey the underlying factors that influent the appearance of body motion. These factors may be related to the individual itself, e.g., ages, gender, emotion, etc. Therefore the body point triplets can be used to analyze the underlying factor inherent to the individual. The body point triplets can also reflect other factors that affect human motion may come from the environments, for instance, uneven grounds, heavy carrying load on the performer, etc. By analyzing the triplet motion, we may also learn such environmental information.

A APPENDIX

A.1 Template matching based on fundamental matrix between views

When pose transition $I_1 \rightarrow I_2$ match to $J_1 \rightarrow J_2$, one can regard the corresponding pairs (e.g. I_1 and J_1) as perspective views of the same object. Since two perspective views of an object are related via epipolar geometry and their associated fundamental matrix [HZ04], a straightforward solution to measure the similarity between $I_{1,2}$ and $J_{1,2}$ would be to check their consistency with the epipolar geometry. Suppose we have $2n$ point correspondences $\mathbf{x}_i \leftrightarrow \mathbf{y}_i$, where $I_1 = \{\mathbf{x}_{1\ldots n}\}$, $I_2 = \{\mathbf{x}_{(n+1),\ldots,2n}\}$, $J_1 = \{\mathbf{y}_{1\ldots n}\}$, $J_2 = \{\mathbf{y}_{(n+1),\ldots,2n}\}$, and n is the number of body points in each image. A classical error function, based on epipolar consistency, can be defined as:

$$\mathcal{E}_s(I_{1,2}, J_{1,2}) = \underset{i=1..2n}{\text{Median}} \left\{ \frac{(\mathbf{y_i^T F x}_i)^2}{(\mathbf{F x}_i)_1^2 + (\mathbf{F x}_i)_2^2 + (\mathbf{F^T y}_i)_1^2 + (\mathbf{F^T y}_i)_2^2} \right\}, \tag{A.1}$$

where $(\cdot)_j$ refers to the j^{th} entry of the vector. The second term in (A.1) is essentially the Sampson cost function [HZ04], which is widely used in estimation methods in multi-view geometry. As discussed in section 2.6.2, for our application $\mathcal{S}_s(\cdot)$ turns out to be less reliable than the proposed matching score function, due to the small number of point correspondences and the noisy localization of body points.

LIST OF REFERENCES

[AC99] JK Aggarwal and Q. Cai. "Human motion analysis: A review." *CVIU*, **73**(3):428–440, 1999.

[AL06] M. Ahmad and S.W. Lee. "HMM-based Human Action Recognition Using Multiview Image Sequences." *ICPR'06*, **1**:263–266, 2006.

[ASB90] K. Arbter, WE Snyder, H. Burkhardt, and G. Hirzinger. "Application of affine-invariant Fourier descriptors to recognitionof 3-D objects." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **12**(7):640–647, 1990.

[AT04] Ankur Agarwal and Bill Triggs. "3D Human Pose from Silhouettes by Relevance Vector Regression." In *CVPR'04*, pp. II 882–888, Washington, June 2004.

[BB81] T. Beardsworth and T. Buckner. "The ability to recognize oneself from a video recording of one's movements without seeing one's body." *Bulletin of the Psychonomic Society*, **18**(1):19–22, 1981.

[BB87] Robert C. Bolles and H. Harlyn Baker. "Epipolar-plane image analysis: a technique for analyzing motion sequences." *Readings in computer vision: issues, problems, principles, and paradigms*, pp. 26–36, 1987.

[BCD02] C. BenAbdelkader, R. Cutler, L. Davis, et al. "Motion-based recognition of people in eigengait space." *International Conference on Automatic Face and Gesture Recognition*, pp. 267–272, 2002.

[BCK78] CD Barclay, JE Cutting, and LT Kozlowski. "Temporal and spatial factors in gait perception that influence gender recognition." *Percept Psychophys*, **23**(2):145–52, 1978.

[BD01a] S.J. Blakemore and J. Decety. "From the perception of action to the understanding of intention." *NATURE REVIEWS NEUROSCIENCE*, **2**(8):561–567, 2001.

[BD01b] AF Bobick and JW Davis. "The recognition of human movement using temporal templates." *TPAMI*, **23**(3):257–267, 2001.

[BD02] G.R. Bradski and J.W. Davis. "Motion segmentation and pose recognition with motion history gradients." *Machine Vision and Applications*, **13**(3):174–184, 2002.

[BGP93] E. Barrett, G. Gheen, and P. Payton. "Representation of Three-Dimensional Object Structure as Cross-Ratios of Determinants of Stereo Image Points." *Proceedings of the Second Joint European-US Workshop on Applications of Invariance in Computer Vision*, pp. 47–68, 1993.

[BGS05] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. "Actions as space-time shapes." In *Proc. ICCV*, volume 2, pp. 1395– 1402, 2005.

[BH00] M. Brand and A. Hertzmann. "Style machines." *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 183–192, 2000.

[BM98] C. Bregler and J. Malik. "Tracking people with twists and exponential maps." *CVPR'98*, pp. 8–15, 1998.

[Car94] S. Carlsson. "The Double Algebra: An Effective Tool for Computing Invariants in Computer Vision." *LECTURE NOTES IN COMPUTER SCIENCE*, pp. 145–145, 1994.

[CBA96] LW Campbell, DA Becker, A. Azarbayejani, AF Bobick, and A. Pentland. "Invariant features for 3-D gesture recognition." *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pp. 157–162, 1996.

[CBP05] Carlo Colombo, Alberto Del Bimbo, and Federico Pernici. "Metric 3D Reconstruction and Texture Acquisition of Surfaces of Revolution from a Single Uncalibrated View." *IEEE Trans. PAMI*, **27**(1):99–114, 2005.

[CCZ00] D. Chi, M. Costa, L. Zhao, and N. Badler. "The EMOTE model for effort and shape." *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 173–182, 2000.

[CF01] R.J. Campbell and P.J. Flynn. "A Survey Of Free-Form Object Representation and Recognition Techniques." *COMPUTER VISION AND IMAGE UNDERSTANDING*, **81**(2):166–210, 2001.

[CPK78] J. Cutting, D. Proffitt, and L. Kozlowski. "A biomechanical invariant for gait perception." *J. of Exp. Psych*, **4**(3):357–372, 1978.

[Dav01] J.W. Davis. "Visual Categorization of Children and Adult Walking Styles." *LECTURE NOTES IN COMPUTER SCIENCE*, pp. 295–300, 2001.

[DG04] J.W. Davis and H. Gao. "An expressive three-mode principal components model for gender recognition." *Journal of Vision*, **4**(5):362–377, 2004.

[DHS01] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*. Citeseer, 2001.

[DK02] J.W. Davis and V.S. Kannappan. "Expressive features for movement exaggeration." *International Conference on Computer Graphics and Interactive Techniques*, pp. 182–182, 2002.

[DLF06] M. Dimitrijevic, V. Lepetit, and P. Fua. "Human body pose detection using Bayesian spatio-temporal templates." *CVIU*, **104**(2-3):127–139, 2006.

[DR05] J. Deutscher and I. Reid. "Articulated Body Motion Capture by Stochastic Search." *IJCV*, **61**(2):185–205, 2005.

[DT02] J. Davis and S. Taylor. "Analysis and Recognition of Walking Movements." *INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION*, **16**:315–318, 2002.

[EBM03] AA Efros, AC Berg, G. Mori, and J. Malik. "Recognizing action at a distance." *ICCV'03*, pp. 726–733, 2003.

[Far99] B. Farnell. "Moving Bodies,Acting Selves." *Annual Review of Anthropology*, **28**:341–373, 1999.

[FCZ98] Andrew W. Fitzgibbon, Geoff Cross, and Andrew Zisserman. "Automatic 3D Model Construction for Turn-Table Sequences." In *SMILE*, pp. 155–170, 1998.

[FH00] P.F. Felzenszwalb and D.P. Huttenlocher. "Efficient matching of pictorial structures." *CVPR'00*, **2**:66–73, 2000.

[FR94] D. Forsyth and C. Rothwell. "Representations of 3D Objects that Incorporate Surface Markings." *Applications of Invariance in Computer Vision: Second Joint European-US Workshop, Ponta Delgada, Azores, Portugal, October 9-14, 1993: Proceedings*, 1994.

[Gav99] DM Gavrila. "Visual analysis of human movement: A survey." *CVIU*, **73**(1):82–98, 1999.

[GBT04] P. Glardon, R. Boulic, and D. Thalmann. "PCA-based walking engine using motion capture data." In *Computer Graphics International*, pp. 292–298, 2004.

[GD96] DM Gavrila and LS Davis. "3-D model-based tracking of humans in action: a multi-view approach." *CVPR'96*, pp. 73–80, 1996.

[GFF96] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti. "Action recognition in the premotor cortex." *Brain*, **119**(2):593, 1996.

[Gol70] A.I. Goldman. *A theory of human action*. Prentice-Hall Englewood Cliffs, NJ, 1970.

[GSS04] A. Gritai, Y. Sheikh, and M. Shah. "On the use of anthropometry in the invariant analysis of human actions." *ICPR'004*, **2**, 2004.

[HHN99] PS Huang, CJ Harris, and MS Nixon. "Recognising humans by gait via parametric canonical space." *Artificial Intelligence in Engineering*, **13**(4):359–366, 1999.

[Hil93] D. Hilbert. *Theory of Algebraic Invariants*. Cambridge University Press, 1993.

[HLF99] N. Howe, M. Leventon, and W. Freeman. "Bayesian reconstruction of 3d human motion from single-camera video." *Neural Information Processing Systems*, **1999**, 1999.

[HS96] G.F. Harris and P.A. Smith. *Human Motion Analysis: Current Applications and Future Directions*. Institute of Electrical & Electronics Engineers (IEEE), 1996.

[HZ04] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[JDL08] I. Junejo, E. Dexter, I. Laptev, and P. Perez. "Cross-view action recognition from temporal self-similarities." In *European Conference on Computer Vision*, volume 12, 2008.

[JOH73] G. JOHANSSON. "Visual perception of biological motion and a model for its analysis." *Perception and Psychophysics*, **14**:201–211, 1973.

[KC77] L.T. Kozlowski and J.E. Cutting. "Recognizing the sex of a walker from a dynamic point-light display." *Perception & Psychophysics*, **21**(6):575–580, 1977.

[KCC03] A. Kale, A.K.R. Chowdhury, and R. Chellappa. "Towards a view invariant gait recognition algorithm." *Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 143–150, 2003.

[KCY03] A. Kale, N. Cuntoor, B. Yegnanarayana, AN Rajagopalan, and R. Chellappa. "Gait Analysis for Human Identification." *LECTURE NOTES IN COMPUTER SCIENCE*, pp. 706–714, 2003.

[KH94] R. Kondepudy and G. Healey. "Use of invariants for recognition of three-dimensional color textures." *Journal of the Optical Society of America A: Optics and Image Science, and Vision*, **11**(11):3037–3049, 1994.

[KM00] L. Kakadiaris and D. Metaxas. "Model-based estimation of 3D human motion." *TPAMI*, **22**(12):1453–1459, 2000.

[LB98] J. Little and J. Boyd. "Recognizing people by their gait: the shape of motion." *Videre: Journal of Computer Vision Research*, **1**(2):1–32, 1998.

[LBP05] I. Laptev, S.J. Belongie, P. Perez, J. Wills, C. universitaire de Beaulieu, and UC San Diego. "Periodic Motion Detection and Segmentation via Approximate Sequence Alignment." *ICCV'05*, **1**:816–823, 2005.

[Lei90] G. Lei. "Recognition of planar objects in 3-D space from single perspectiveviews using cross ratio." *Robotics and Automation, IEEE Transactions on*, **6**(4):432–437, 1990.

[LG02] L. Lee and WEL Grimson. "Gait analysis for recognition and classification." *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pp. 148–155, 2002.

[MDK64] M. MURRAY, A.B. DROUGHT, and R.C. KORY. "Walking Patterns of Normal Men." *The Journal of Bone and Joint Surgery*, **46**(2):335, 1964.

[MG00] T.B. Moeslund and E. Granum. "3D human pose estimation using 2D-data and an alternative phase space representation." *Workshop on Human Modeling, Analysis and Synthesis at CVPR*, **16**, 2000.

[MG01] T.B. Moeslund and E. Granum. "A survey of computer vision-based human motion capture." *CVIU*, **81**(3):231–268, 2001.

[MM02] G. Mori and J. Malik. "Estimating human body configurations using shape context matching." *European Conference on Computer Vision*, **3**:666–680, 2002.

[MM06] G. Mori and J. Malik. "Recovering 3 D Human Body Configurations Using Shape Contexts." *IEEE Trans. PAMI*, **28**(7):1052–1062, 2006.

[MUT91] Y. Moses, S. Ullman, Massachusetts Institute of Technology, and Artificial Intelligence Laboratory. *Limitations of Non Model-Based Recognition Schemes.* Springer, 1991.

[MW97] Q. Minh and W. Wageeh. "Wavelet-Based Affine Invariant Representation: A Tool for Recognizing Planar Objects in 3D Space." *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, pp. 846–857, 1997.

[OMB06] E.J. Ong, A.S. Micilotta, R. Bowden, and A. Hilton. "Viewpoint invariant exemplar-based 3D human tracking." *CVIU*, **104**(2-3):178–189, 2006.

[PC03] V. Parameswaran and R. Chellappa. "View invariants for human action recognition." *CVPR'03*, **2**, 2003.

[PLR02] F.E. Pollick, V. Lestou, J. Ryu, and S.B. Cho. "Estimating the efficiency of recognizing gender and affect from biological motion." *Vision Research*, **42**(20):2345–2355, 2002.

[PRC99] V. Pavlovic, J.M. Rehg, T.J. Cham, and K.P. Murphy. "A Dynamic Bayesian Network Approach to Figure Tracking using Learned Dynamic Models." *ICCV(1)*, pp. 94–101, 1999.

[Pri97] W. Prinz. "Perception and Action Planning." *European Journal of Cognitive Psychology*, **9**(2):129–154, 1997.

[Rab89] LR Rabiner. "A tutorial on hidden Markov models and selected applications inspeech recognition." *Proceedings of the IEEE*, **77**(2):257–286, 1989.

[RF03] D. Ramanan and DA Forsyth. "Finding and tracking people from the bottom up." *CVPR'03*, **2**, 2003.

[RK95] J. Rehg and T. Kanade. "Model-based tracking of self-occluding articulated objects." *ICCV*, pp. 612–617, 1995.

[RMR04] T.J. Roberts, S.J. McKenna, and I.W. Ricketts. "Human pose estimation using learnt probabilistic region similarities and partial configurations." *ECCV*, **4**:291–303, 2004.

[RPL05] A. Rav-Acha, Y. Pritch, D. Lischinski, and S. Peleg. "Dynamosaics: video mosaics with non-chronological time." *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, **1**, 2005.

[RS00a] R. Rosales and S. Sclaroff. "Inferring body pose without tracking body parts." *PROC IEEE COMPUT SOC CONF COMPUT VISION PATTERN RECOGNIT*, **2**:721–727, 2000.

[RS00b] R. Rosales and S. Sclaroff. "Specialized mappings and the estimation of human body pose from asingle image." *Human Motion, 2000. Proceedings. Workshop on*, pp. 19–24, 2000.

[RYS02] C. Rao, A. Yilmaz, and M. Shah. "View-Invariant Representation and Recognition of Actions." *IJCV*, **50**(2):203–226, 2002.

[SBF00] H. Sidenbladh, M.J. Black, and D.J. Fleet. "Stochastic tracking of 3D human figures using 2D image motion." *ECCV*, **2**:702–718, 2000.

[SCK96] J. Subrahmonia, DB Cooper, and D. Keren. "Practical reliable Bayesian recognition of 2D and 3D objects usingimplicit polynomials and algebraic invariants." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **18**(5):505–519, 1996.

[SD97] S.M. Seitz and C.R. Dyer. "View-Invariant Analysis of Cyclic Motion." *International Journal of Computer Vision*, **25**(3):231–251, 1997.

[SKL05] Cristian Sminchisescu, Atul Kanaujia, Zhiguo Li, and Dimitris N. Metaxas. "Discriminative Density Propagation for 3D Human Motion Estimation." In *CVPR*, pp. 390–397, 2005.

[SLC04] C. Schuldt, I. Laptev, and B. Caputo. "Recognizing human actions: a local SVM approach." *ICPR'04*, **3**, 2004.

[SLD01] G. Shakhnarovich, L. Lee, and T. Darrell. "Integrated Face and Gait Recognition from Multiple Views." *IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION*, **1**, 2001.

[SLW98] C.W. Sul, K.C. Lee, and K. Wohn. "Virtual Stage: a location-based karaoke system." *Multimedia, IEEE*, **5**(2):42–52, 1998.

[SNS02] WI Schöllhorn, BM Nigg, DJ Stefanyshyn, and W. Liu. "Identification of individual walking patterns using time discrete and time continuous data sets." *Gait & Posture*, **15**(2):180–186, 2002.

[SNV99] S.V. Stevenage, M.S. Nixon, and K. Vince. "Visual analysis of gait as a cue to identity." *Applied Cognitive Psychology*, **13**(6):513–526, 1999.

[SRC03] A. Sundaresan, A. RoyChowdhury, and R. Chellappa. "A hidden Markov model based framework for recognition of humans from gait sequences." *Image Processing, 2003. Proceedings. 2003 International Conference on*, **2**, 2003.

[SS05] Y.S. Sheikh and M.M. Shah. "Exploring the Space of a Human Action." *ICCV'05*, **1**, 2005.

[ST93] G. Sapiro and A. Tannenbaum. "Affine invariant scale-space." *International Journal of Computer Vision*, **11**(1):25–44, 1993.

[ST01] C. Sminchisescu and B. Triggs. "Covariance scaled sampling for monocular 3D body tracking." *CVPR'01*, **1**:447–454, 2001.

[ST02] C. Sminchisescu and A. Telea. "Human pose estimation from silhouettes. a consistent approach using distance level sets." *WSCG International Conference on Computer Graphics, Visualization and Computer Vision*, 2002.

[ST03] C. Sminchisescu and B. Triggs. "Kinematic jump processes for monocular 3D human tracking." *CVPR'03*, **1**, 2003.

[SVS01] T. Syeda-Mahmood, A. Vasilescu, S. Sethi, I.B.M.A.R. Center, and CA San Jose. "Recognizing action events from multiple viewpoints." *Detection and Recognition of Events in Video, 2001. Proceedings. IEEE Workshop on*, pp. 64–72, 2001.

[Tay00] CJ Taylor. "Reconstruction of articulated objects from point correspondences ina single uncalibrated image." *CVPR'00*, **1**, 2000.

[TC03] D. Tolliver and R.T. Collins. "Gait Shape Estimation for Identification." *LECTURE NOTES IN COMPUTER SCIENCE*, pp. 734–742, 2003.

[TP91] M. Turk and A. Pentland. "Face recognition using eigenfaces." In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 591, 1991.

[Tro02] N.F. Troje. "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns." *Journal of Vision*, **2**(5):371–387, 2002.

[TW] J.B. Tenenbaum and T. William. "Freeman. Separating style and content." *Advances in Neural Information Processing Systems*, **9**:662–668.

[UAT95] M. Unuma, K. Anjyo, and R. Takeuchi. "Fourier principles for emotion-based human figure animation." *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pp. 91–96, 1995.

[Vas02] M. Vasilescu. "Human Motion Signatures: Analysis, Synthesis, Recognition." *INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION*, **16**:456–460, 2002.

[VT02] M.A.O. Vasilescu and D. Terzopoulos. "Multilinear Analysis of Image Ensembles: TensorFaces." *LECTURE NOTES IN COMPUTER SCIENCE*, pp. 447–460, 2002.

[Wan06] L. Wang. "Abnormal Walking Gait Analysis Using Silhouette-Masked Flow Histograms." *ICPR'06*, **3**:473–476, 2006.

[WB99] A.D. Wilson and A.F. Bobick. "Parametric Hidden Markov Models for Gesture Recognition." *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, pp. 884–900, 1999.

[Wei93] I. Weiss. "Geometric invariants and object recognition." *International Journal of Computer Vision*, **10**(3):207–231, 1993.

[Wei94] I. Weiss. "Invariants for Recovering Shape from Shading." *LECTURE NOTES IN COMPUTER SCIENCE*, pp. 185–185, 1994.

[WRB06] Daniel Weinland, Remi Ronfard, and Edmond Boyer. "Free viewpoint action recognition using motion history volumes." *CVIU*, **104**(2-3):249–257, 2006.

[WS07] L. Wang and D. Suter. "Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model." In *CVPR'07*, pp. 1–8, 2007.

[WTN03] L. Wang, T. Tan, H. Ning, and W. Hu. "Silhouette analysis-based gait recognition for human identification." *IEEE Trans. PAMI*, **25**(12):1505–1518, 2003.

[YS05] A. Yilmaz and M. Shah. "Actions sketch: a novel action representation." *CVPR'05*, **1**, 2005.

[YS06] A. Yilmaz and M. Shah. "Matching actions in presence of camera motion." *CVIU*, **104**(2-3):221–231, 2006.

[Zat02] V.M. Zatsiorsky. *Kinetics of Human Motion*. Human Kinetics, 2002.

[ZCL04] J. Zhang, R. Collins, and Y. Liu. "Representation and matching of articulated shapes." *CVPR'04*, **2**, 2004.

[ZL01] Z. Zhang and C. Loop. "Estimating the fundamental matrix by transforming image points in projective space." *CVIU*, **82**(2):174–180, 2001.

[ZN94] M. Zerroug and R. Nevatia. "Using Invariance and Quasi-Invariance for the Segmentation and Recovery of Curved Objects." *LECTURE NOTES IN COMPUTER SCIENCE*, pp. 317–317, 1994.

[ZT98] Z.Y. Zhang and H.T. Tsui. "3D Reconstruction from a Single View of an Object and Its Image in a Plane Mirror." *INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION*, **14**:1174–1176, 1998.

[ZXG06] G. Zhu, C. Xu, W. Gao, and Q. Huang. "Action Recognition in Broadcast Tennis Video Using Optical Flow and Support Vector Machine." *LNCS*, **3979**:89, 2006.