MULTI-VIEW GEOMETRIC CONSTRAINTS FOR HUMAN ACTION RECOGNITION AND TRACKING

by

ALEXEI GRITAI
B.S. Moscow Institute of Electronic Technology, 1992
M.S. Moscow Institute of Electronic Technology, 1994
M.S. University of Central Florida, 2003

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

Spring Term
2007

Major Professor: Mubarak Shah

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

Human actions are the essence of a human life and a natural product of the human mind. Analysis of human activities by a machine has attracted the attention of many researchers. This analysis is very important in a variety of domains including surveillance, video retrieval, human-computer interaction, athlete performance investigation, etc. This dissertation makes three major contributions to automatic analysis of human actions. First, we conjecture that the relationship between body joints of two actors in the same posture can be described by a 3D rigid transformation. This transformation simultaneously captures different poses and various sizes and proportions. As a consequence of this conjecture, we show that there exists a fundamental matrix between the imaged positions of the body joints of two actors, if they are in the same posture. Second, we propose a novel projection model for cameras moving at a constant velocity in 3D space, *Galilean* cameras, and derive the Galilean fundamental matrix and apply it to human action recognition. Third, we propose a novel use for the invariant ratio of areas under an affine transformation and utilizing the epipolar geometry between two cameras for 2D model-based tracking of human body joints.

In the first part of the thesis, we propose an approach to match human actions using semantic correspondences between human bodies. These correspondences are used to provide geometric constraints between multiple anatomical landmarks (e.g. hands, shoulders, and feet) to match actions observed from different viewpoints and performed at different rates by actors of differing anthropometric proportions. The fact that the human body has approximate anthropometric proportion allows for innovative use of the machinery of epipolar geometry to provide constraints for analyzing actions performed by people of different anthropometric sizes, while ensuring that changes in viewpoint do not affect matching. A novel measure in terms of rank of matrix constructed only from image measurements of the locations of anatomical landmarks is proposed to ensure that similar actions are accurately recognized. Finally, we describe how dynamic time warping can be used in conjunction with the proposed measure to match actions in the presence of nonlinear time warps. We demonstrate the versatility of our algorithm in a number of challenging sequences and applications including action synchronization , odd one out, following the leader, analyzing periodicity etc.

Next, we extend the conventional model of image projection to video captured by a camera moving at constant velocity. We term such moving camera Galilean camera. To that end, we derive the spacetime projection and develop the corresponding epipolar geometry between two Galilean cameras. Both perspective imaging and linear pushbroom imaging form specializations of the proposed model and we show how six different "fundamental" matrices including the classic fundamental matrix, the Linear Pushbroom (LP) fundamental matrix, and a fundamental matrix relating Epipolar Plane Images (EPIs) are related and can be directly recovered from a Galilean fundamental matrix. We provide linear algorithms for estimating the parameters of the the mapping between videos in the case of planar scenes. For applying fundamental matrix between Galilean cameras to human action recognition, we propose a measure that has two important properties. First property makes it possible to recognize similar actions, if their execution rates are linearly related. Second property allows recognizing actions in video captured by Galilean cameras. Thus, the proposed algorithm guarantees that actions can be correctly matched despite changes in view, execution rate, anthropometric proportions of the actor, and even if the camera moves with constant velocity.

Finally, we also propose a novel 2D model based approach for tracking human body parts during articulated motion. The human body is modeled as a 2D stick figure of thirteen body joints and an action is considered as a sequence of these stick figures. Given the locations of these joints in every frame of a model video and the first frame of a test video, the joint locations are automatically estimated throughout the test video using two geometric constraints. First, invariance of the ratio of areas under an affine transformation is used for initial estimation of the joint locations in the test video. Second, the epipolar geometry between the two cameras is used to refine these estimates. Using these estimated joint locations, the tracking algorithm determines the exact location of each landmark in the test video using the foreground silhouettes. The novelty of the proposed approach lies in the geometric formulation of human action models, the combination of the two geometric constraints for body joints prediction, and the handling of deviations in anthropometry of individuals, viewpoints, execution rate, and style of performing action. The proposed approach does not require extensive training and can easily adapt to a wide variety of articulated actions.

This dissertation is dedicated to my Dad, my best friend

ACKNOWLEDGMENTS

I wish to express my gratitude to everyone who contributed to making my Ph.D. dissertation a reality. I must single out the Director of the Computer Vision Lab, Agere Chair Professor Dr. Mubarak Shah, who supported this dissertation during the years. I owe a special debt to my advisor, Dr. Mubarak Shah. Thank you for your guidance on this research topic, and on research and life in general. You are simply the best advisor, period. I thank Dr. Shah for teaching me all that I know about Computer Vision, for being stubborn, disciplined and organized with my research, and for the many lessons on writing.

I also want to thank a former student of Dr. Shah and now a scientist researcher in Sarnoff Corporation, Dr. Cen Rao , for guiding my entrance into Human Action Recognition field and for his contribution to this area that gave me a solid fundament to built my own theory. I am again grateful to Agere Chair Professor Dr. Mubarak Shah for giving me an opportunity to collaborate with Dr. Cen Rao.

I would like to gratefully acknowledge another former student of Dr. Shah, who is now a Postdoc Fellow in Carnegie Mellon University, Dr. Yaser Sheikh, for his collaboration, insightful comments, and helpful advice in my research. I am very grateful to Dr. Yaser Sheikh for his guidance in Multiple View Geometry, priceless help, and lessons on writing.

This research has also benefited tremendously from all of the members of the Computer Vision Lab at UCF. Special thanks to Arslan Basharat, Mikel Rodriguez, and Eric Leach for countless hours spent editing and discussing the dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

"All human actions have one or more of these seven causes: chance, nature, compulsion, habit, reason, passion, and desire".

Aristotle (384 BC - 322 BC)

Human actions are the essence of a human life and a natural product of the human mind. Aristotle, Confucius, Demosthenes were first among many other philosophers and scientists who tried in their immortal treatises to understand the role of people's activities from different aspects. In his landmark treatise titled *Human Actions*, [Mis66], Ludwig Von Mises opens his first chapter with the statement, *"Human action is purposeful behavior"*. He states that actions ostensibly reflect the actor's intention, conscious, or unconscious. It is not surprising, therefore, that visual perception of actions is a critical cognitive function for interpreting the intention of an observed actor and for understanding the observer's environment. As social entities, the ability to interpret and predict the behavior of others is fundamental to normal human functioning. In fact, there is a growing body of evidence that humans might actually understand the actions of another individual in terms of the same neural code that they use to produce the same action themselves, [DG99]. It is for these reasons that the analysis of human actions is a subject of interest in a number of scientific communities such as philosophy [Gol70], developmental psychology [Pri97], economics [Mis66], and recently in cognitive neuroscience [VR96, BD01]. It is also why developing algorithms for action understanding must figure prominently in the pursuit of both machine intelligence and robotics.

Until the middle of the 20^{th} century people explored the role of human activities in a non-rigorous manner, but in 1956 all accumulated knowledge was organized into what became a new area in Computer Science. The name of that area was officially declared as *Artificial Intelligence (AI)*. Once established as a field in a scientific community, researchers explored the various aspects of human activity. In the 1970's the birth of an area within AI, *Computer Vision (CV)*, a new era in understanding of human actions began. With increasing computation power, computer vision scientists contributed many breakthroughs in the analysis of human actions. However, many problems in the analysis of human actions remain unsolved, and because of this, the area of human actions remains an object of constant study and attention.

The major steps in human motion analysis are *human detection*, *tracking*, and *human action recognition*. Constant attention to human motion analysis has extended each step intro a broad area of research. Each of these steps include 2D and 3D approaches based on the use of either multiple or monocular cameras. The human detection phase includes methods ranging from the detection of people in images to determining the 3D pose estimation of a human body. The means for tracking people vary from tracking human body centroids to the estimation of either 2D or 3D positions of landmarks on a human body given a sequence of images. Recognizing and understanding a human's behavior in a video is one of the ultimate goals of human motion analysis, and algorithms oriented on this task constitute the class known as human action recognition. Action recognition algorithms are high level approaches and depend on the success of human detection and tracking results. This dissertation proposes a series of methods to solve problems belonging to the classes of *recognition* and *tracking* in a sequence of 2D images.

1.1 Human Action Recognition

Developing algorithms to recognize human actions has proven to be a significant challenge since it is a problem that combines the uncertainty associated with computational vision along with unpredictable human behavior. Even without these two sources of variability, the human body has no less than 244 degrees of freedom ([Zat02]) and modeling the dynamics of an object with such non-rigidity is an extremely difficult task. Further compounding the problem, recent research into anthropology has revealed that body dynamics are far more complicated than was earlier thought. It has been found to be affected by age, ethnicity, class, family tradition, gender, skill, circumstance, and choice [Far99]. Human actions are not merely functions of joint angles and anatomical landmark positions, they affected by the psychological state as well as the cultural background of the actor. Thus, the sheer range and complexity of human actions makes developing action recognition algorithms a daunting task.

To develop computer algorithms for action analysis, it is important to identify properties that are expected to vary according to a set of transformations with each observation of an action, but which should not affect recognition:

Viewpoint: Except in certain specific applications, it is unreasonable, in general, to assume that the viewpoint from which actions are observed would remain constant across different observations of that action. Thus, it is important that algorithms for action recognition exhibit stability in recognition despite large changes in viewpoint. The relationship of action recognition to object recognition was observed by Rao and Shah in [RS01], and developed further by Parameswaran and Chellappa in [PC03], [PC02], Gritai *et al.* in [GSS04], and by Yilmaz and Shah in [YS05a]. In these papers, the importance of view invariant recognition has been stressed, highlighting the fact that, as in object recognition [Ver92], the vantage point of the camera should not affect recognition. The projective and affine geometry of multiple views is well-understood, see [HZ00], and various invariants have been proposed. There has also been some discussion of viewpoint variance and invariance in cognitive neuroscience in the context of both object and action recognition, [Ver92, DV99]. In the proposed approach, accurate matching in the presence of varying viewpoint is a central problem which we address by using geometric relationships between the two observed executions of an action.

Anthropometry: In general, an action can be executed irrespective of the size or gender of the actor. It is therefore important that action recognition be unaffected by "anthropometric transformations". Unfortunately, since anthropometric transformations do not obey any known laws, characterizing invariants formally is impossible. However, empirical studies have shown that these transformations are not *arbitrary* (see [RC82]). The study of human proportions has a great tradition in science, from the 'Golden Sections' of ancient China, India, Egypt and Greece down to renaissance thinkers like Leornardo Da Vinci (the Vitruvian Man) and Albrecht Durer, with modern day applications in Ergonomics and human performance engineering. We provide a functional definition of anthropometric transforms making implicit use of the 'laws' governing human body proportions to provide geometric constraints for matching. Instead of using a single point representation, we explore the use of several points on the actor for action recognition, and use geometric constraints with respect to two *actors* performing the action instead of two camera *views*. This innovative use of geometry allows two interesting results for the recognition of actions. The first result provides a constraint to measure the dissimilarity of the posture of two actors viewed in two images. The second result extends this first constraint to globally measure dissimilarity between two actions.

Execution Rate: With rare exceptions such as synchronized dancing or army drills, actions are rarely executed at a precise rate. Furthermore, the cause of temporal variability can be two fold, caused by the actor or possibly by differing camera frame-rates. It is desirable, therefore, that action recognition algorithms remain unaffected by some set of temporal transformations. To handle general nonlinear temporal transformations, we utilize Dynamic Time Warping (DTW) for matching, which ensures only that the temporal order is preserved.

1.2 Galilean Matrix and its Application in Action Recognition

Hitherto an enormous quantity of human action recognition algorithms have been designed for stationary cameras. In this work, we propose a method that is able to recognize human actions observed by cameras moving with a constant velocity in any direction. Similarly to the fundamental matrix governing the relation between two images captured by stationary cameras observing the same scene, there exists a matrix that encapsulates the coherence between images from the same scene captured by cameras moving with a constant speed in any direction.

In order to design such an algorithm, we propose a new spacetime projection for uniformly moving cameras. We generalize conventional imaging projection to model video captured by a camera moving with constant velocity. To that end, we introduce the concept of spacetime projection and develop the corresponding epipolar geometry between two such cameras. We provide a linear algorithm for estimating the parameters of the resulting Galilean "fundamental" matrix. Both perspective imaging and linear pushbroom imaging form a specialization of the proposed model, and we show how different fundamental matrices including the original fundamental matrix, the Linear Pushbroom fundamental matrix, a fundamental matrix relating Epipolar Plane Images, and fundamental matrices relating all permutations between these images can be directly recovered from the fundamental matrix of uniform motion.

Applying Galilean fundamental matrix to human action recognition, we propose a measure that has two important properties. First property ensures the possibility to recognize similar actions, if their execution rate are linearly coherent. Second property allows to recognize actions in video captured by Galilean cameras. Thus, the proposed algorithm guarantees that actions can be correctly matched despite changes in view, execution rate, anthropometric proportions of the actor, and even if the camera moves with constant velocity in 3D space.

1.3 Human Joint Tracking

Analysis of human motion and activities by a machine has attracted the attention of many researchers. This analysis is very important in a variety of domains including surveillance, human-computer interaction, action recognition, athlete performance investigation, etc. The details of the available inputs and the required output of motion analysis depends on the domain. Detection and tracking of different body parts (arms, legs, torso, etc.) or landmark points (elbows, knees, shoulders, etc.) provides the low level information for a variety of applications in human motion analysis. This forms the foundation for a variety of higher level approaches, e.g., human-computer interaction, action recognition, etc. The success of these applications strongly relies on the accuracy of body joint detection and tracking. Issues like self-occlusion, articulated motion, variation in execution rates, and non-discriminative appearance make the problem of joint tracking very challenging. The proposed work performs the low level analysis of human motion by predicting and tracking the location of 13 body joints by using a 2D model of the action being performed.

We address the problem of detecting and tracking human body parts (head, torso, arms, and legs) in a test video when the same action has been performed in a model video. The locations of the 13 joints are assumed to be available in the model video and in the first frame of the test video. The action in the test video may be performed with some variation in the viewpoint, anthropometry of individuals, and execution rate and style. The epipolar and affine constraints are applied to predict the locations of these joints in a test video. These predicted points are then used along with foreground silhouettes of the human body to detect and track the joint locations.

The 2D action model is defined as a sequence of 2D stick figures that are composed of thirteen landmark points on the human body. The motivation of our prediction approach lies in the underlying geometrical similarity between the structures of the model and test actions. Two geometrical constraints are applied to *map* the model action onto the test action. This mapping is nontrivial due to four main factors that include variations in the viewpoints of cameras, anthropometry of the two individuals performing the action, action execution rate, and the style in which they perform the action. The first two factors are handled in the prediction phase while the last two are addressed in the tracking phase. The first geometrical constraint in the prediction phase is based on the invariance of a ratio of areas under an affine transformation [HZ00]. Under this constraint, the geometry of the action model is used to recover an initial estimate of the predicted joint locations in the test video. These estimates are improved by using the epipolar constraint, where the fundamental matrix between model and test views is used to determine the final set of joint predictions. The combination of these two constraints helps in handling significant variations in anthropometry of individuals and quasi variation between the viewpoints. The appendix presents an analysis of sensitivity to the degree of viewpoint variation. The prediction phase is followed by the tracking phase in which the variations in action execution (both rate and style) are handled and the features from the test video are used to detect and track all 13 body joints. Our tracking approach does not make any assumptions about the rate of action in the test video, and uses the set of predictions to handle occlusion. Out of the four variations between the model and test actions, the proposed approach is invariant to variations in two (anthropometry and execution rate) of them, while quasi invariant to variations in the other two (viewpoint and execution style). The main contributions of this work include the unique geometrical formulation of human action models, the combination of the two geometric constraints for prediction of the joint locations in the test video, and the handling of the aforementioned variations between model and test actions through prediction and tracking phases.

1.4 Organization of the Dissertation

In Chapter 2, we discuss the related work for human action recognition, different types of fundamental matrices and its application to human action recognition, and human joint tracking. In Section 2.1, we observe feature-based approaches and direct approaches and discuss on their advantages and disadvantages. We explain why the majority of action recognition algorithms suffer from a variation in the camera's viewpoint. In Section 2.2, we discuss the fundamental matrices between two images captured by a stationary pinhole camera, a moving stereo pair, a moving X-Slits camera, and a moving pinhole camera. In Section 2.3, we describe some major directions that have been proposed in recent years for tracking human joints. The general tendency in these approaches has been to increase the accuracy of human joint tracking without increasing complexity.

In Chapter 3, we discuss an action representation, and propose a multiple trajectory representation. We conjecture that there is a projective transformation in 3D space between people. The conjecture implies *action* and *postural* constraints, and allows us to explicitly state view, anthropometric, and temporal transformations. We prove and empirically demonstrate that our method is robust to all important transformations, and can be used in different applications.

Chapter 4 is dedicated to the discussion of a spacetime projection model, a homography, and a fundamental constraint between *Galilean* cameras. From the Galilean fundamental matrix, we derive fundamental constraints between six different pairs of images: perspective and perspective, perspective and pushbroom, perspective and epipolar plane, pushbroom and pushbroom, pushbroom and epipolar plane, epipolar plane and epipolar plane images. Similar to the application of the standard fundamental matrix for action recognition, the Galilean fundamental matrix is applied to human action recognition, and as a consequence, human actions can be recognized in video captured by uniformly moving cameras.

In Chapter 5, we present a novel method for the 2D motion reconstruction of human body joints using human action exemplars and multi-view geometric constraints. The innovative use of these constraints significantly reduces the search space for tracking joints, and hence, reduces the complexity of the algorithm without a loss in tracking accuracy. In addition, the proposed method does not require extensive training, and can even be used with a single action model. The application of an affine and epipolar line constraints is demonstrated on 2D human joint tracking in monocular video captured with a significant change in camera viewpoint with respect to a video containing the model action.

In Chapter 6, we summarize the contributions and results of the presented approaches for action recognition, estimation multi-view geometric constraint between uniformly moving cameras (and its application for action recognition), and 2D human joint tracking in monocular video.

CHAPTER 2
RELATED WORK

Research on human action recognition through computer vision began in the late seventies with the earliest work probably being the PhD thesis of Herman, [Her79]. This work used a static representation of a stick figure in a single image to analyze different postures of a person. The importance of recognizing and understanding human actions in sequence of images was almost immediately realized and materialized in a series of papers in the the early eighties, [RB80], [Aki84], [Ras80]. Since then, a large body of literature has accumulated with studies different approaches to detect, track, reconstruct, and recognize human motion. Surveys in this area have been regularly published, including Aggarwal *et al.*, [JS94] in 1994, Cedras and Shah, [CS95], in 1995, Ju [JBY96] in 1996, Aggarwal and Cai, [AC99], Gavrila, [Gav99], in 1999, Moeslund and Granum, [MG01], in 2001, Buxton, [Bux03], Wang, [LT03], in 2003, Aggarwal and Park, [AP04], in 2004, and finally, Moeslund *et al.*, [MHK06] in 2006. Under Gavrila's taxonomy of human motion analysis, methods can be roughly classified as image-based approaches or 3D approaches, i.e., methods that try to recover and analyze 3D information of human postures and dynamics, as well as those that perform recognition directly from 2D image measurements.

2.1 Human Action Recognition

Typically in 3D approaches, models of human body and human motion are used and a projection of the model in a particular posture is then compared with each frame of the input video to recognize the action. The advantage of these approaches is that since a 3D model is explicitly used these methods are inherently view invariant. However, they are usually computationally quite expensive, [Hog84] and 3D recovery of the articulated objects is still a difficult problem. As a result, 3D approaches are therefore usually limited in some specific applications, such as athletic analysis and sign language recognition [CBA96, DS94].

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

A whole slew of different features have been proposed and used. To recognize the temporal textures, the statistical features of optical flow such as mean flow magnitude, standard deviation, the positive and negative curl, and divergence, are used in [PN94]. Other features to recognize human activities include region-based [DB97, NA94, PN94, AS98, LG05], temporal trajectory-based [NOM98, YA98, RS01, GS89], part-based [BY95, BHB00, JBY96], or a combination of these [BJ98, HHD00]. The approaches work based on features capturing either 2D shape or motion information. Usually, the recognition system involves some dissimilarity or similarity measurement between the activities and the models, such as the shape of the silhouettes, the trajectories of the moving hands, or the point clouds from the body parts. Hidden Markov models have also been a popular tool in using these features for recognition following its success in speech recognition [YXC97]. The earliest papers included work by Pentland *et al* [SP96] and Yamato *et al* [YOI95]. More sophisticated models, such as Coupled Hidden Markov Model (CHHM)[ORP99], Variable Length Markov Model (VLMM)[AH01], Layered Hidden Markov Model (LHMM) [OHG02], stochastic context free grammar (SCFG) [BI98], and Hierarchical Hidden markov model (HHMM) [NB05, SN] have been proposed for efficiently representing and recognizing activities from one or more persons. However these method require training data, and generally lack the capability of explaining the actions semantically.

Most recently, approaches loosely applying the paradigm of 'direct' methods proposed by Horn and Weldon in [HW88] which utilize the spatiotemporal information directly for motion analysis, have started to appear. The difference from feature based approaches is that image measurements are *directly* used for recognition. An approach based on the statistical features of spatiotemporal gradient direction is used for classifying human activities, e.g., walking, running, and jumping [CI00]. In [ZI01], an action recognition system is proposed by matching the histogram of the optical flow generated by different actions. This approach is extended in [SI05], so that the spatiotemporal volumes of actions are exploited and a correlation measure is computed for recognizing the same action from different video. The spatiotemporal information of actions is further used for detecting irregularities in images and in video [BI05]. In this work, a statistical framework is proposed for matching the patches containing actions in the video. In [YH05], Ke *at al* proposed using boosted classifiers to detect action events from the video from simple spatiotemporal filters. In [BGS05], the silhouettes of the moving subjects are used in addition to the spatiotemporal information of the pixels. The method utilizes properties of the solution to the Poisson equation to extract spacetime features such as local spacetime saliency, action dynamics, shape structure and orientation. These features are used for action recognition, detection and clustering.

The fundamental drawback of using such 2D image-based approaches (direct approaches in particular) is that they are viewpoint dependent. An intermediate category of approaches, including this paper, use image measurements, but exploit 3D constraint by exploiting the geometry of multiple views. Seitz and Dyer [SD97] used view-invariant measurement to find the repeating pose of walking people and the reoccurrence of position of turning points. Laptev [LBP05] proposed using spatiotemporal 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. Parameswaran and Chellappa 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 [PCar]. The multiple trajectories from the joints of a person are recorded, the pose during the action is matched with a canonical body pose, and the matching coefficients are used for representing the action, and the temporal variance of the actions is compensated using DTW. Finally, the actions are matched by comparing the coefficients of the actions.

2.2 Fundamental Matrix and its use in Human Action Recognition

In 1992 Luong [FL01] and Hartley [HZ00] proposed a fundamental matrix between two perspective images of the same scene captured from distinct viewpoints. The fundamental matrix is the algebraic form of the intrinsic geometry between two views. Within just a few years applications for the fundamental matrix were found in 3D scene reconstruction, stereo camera applications, image alignment, video synchronization, etc. In [HZ00], Hartley and Zisserman summarized the work related to the fundamental matrix between multiple views of the same static scene. Knowing the significance of the fundamental matrix, researchers tried to extend its application to other types of images and scenes. By relaxing constraints on static cameras and scenes, researchers explored new cases of the geometric coherence.

In [AS00], Avidan and Shashua considered the case where objects can move freely along lines or conics in 3D, and there is no constraint on the camera motion. In this case, the 3D motion trajectories can be recovered from sequence of images.

Similarly, Han and Kanade in [HK00] constrained the object's motion along lines in 3D, but allow a scene to contain both moving and stationary objects. By recovering the camera motion, static and moving objects can be automatically segmented out for direct application to 3D scene reconstruction. By relaxing the constraint on the constant speed of moving objects and imposing planar motion constraint, Bartoli in [Bar03] proposed a multi-view C-tensor that is similar to the multi-focal tensor.

In [Stu02], Sturm considered a moving stereo system, observing rigid objects that move arbitrarily along the plane. At each time instant the moving stereo system gives a 3D view of a scene. A tensor for matching different 3D views of a scene was proposed. The proposed tensor is analogous to the fundamental matrix between two perspective images.

Considering satellite imagery (static scene and camera moving linearly in 3D), in [GH97] Gupta and Hartley developed the geometric relation between such images. Authors proposed to model the satellite image by a *Linear Pushbroom* image, which is a perspective along the direction of motion and orthographic along direction perpendicular to the motion. As a result, a novel fundamental matrix relating linear pushbroom images was proposed.

So far, researchers have considered images as these obtained from a pinhole camera. In the 19th century, Ducos du Hauron designed the X-Slits camera, that was forgotten until recently. In this model, the centers of horizontal and vertical projections lie in different locations on the camera's optical axis [DF02]. In [DF02], Weinshall *et al.* proposed a method to generate new type of images from a bi-centric camera (X-Slits camera). It was demonstrated that perspective and linear pushbroom images can be simply captured by a bi-centric camera. Images from X-Slits cameras allows for generating virtual positions in a scene that can not be always simulated from images captured by the pinhole camera. In [AW03], Zomet *et al.* revised the X-Slits camera, proposed the nonperspective projection model of the X-Slits camera, and discussed its properties. Compared to pinhole cameras, the X-Slits camera has strong advantages in image rendering. When compared to traditional mosaicing, X-Slits images can be shown to be closer to perspective images than linear pushbroom images [AW03]. In [DW02], Feldman *et al.* developed the epipolar geometry of the X-

Slits projection model. The epipolar geometry of X-Slits cameras resembles the pinhole epipolar geometry, but has unique proerties.

In [SS06], Khan *et al.* developed the ortho-perspective fundamental matrix between perspective and linear pushbroom images captured by a traditional pinhole camera. Once the ortho-perspective fundamental matrix is computed, the camera location corresponding to the perspective image can be recovered by finding its epipole in route panorama image.

Our work is related to that of Wolf and Shashua in [WZ02], where they investigate higher-dimensional mappings between k-spaces and 2-spaces that arise from different problem instances for $3 \leq k \leq 6$. They provide six problem definitions describing various configurations of planar and non-planar points moving in straight lines viewed by general cameras.

Compared to previous approaches, we describe a spacetime projection model for a Galilean camera and propose a mapping function between the videos of two Galilean cameras when the scene is planar. We present the epipolar geometry for this case and describe a normalized linear algorithm for estimating the parameters of the "fundamental" matrix relating Galilean cameras. We show how the original fundamental matrix, the LP fundamental matrix, the ortho-perspective fundamental matrix, and three other, as of yet unknown, fundamental matrices can be directly recovered from this Galilean matrix.

In recent years, the use of the standard fundamental matrix in action recognition has achieved great attention because its allows us to recognize human actions under a significant variation in the camera viewpoint. In [RGS03], Rao *et al.* applied the fundamental matrix to 2D human motion synchronization. In [YS05b, YS05a], Yilmaz and Shah applied a multi-view geometric constraint to human action recognition, and similarly, in [YS05c], they proposed a temporal fundamental matrix to recognize actions captured by arbitrarily moving cameras. This work generalizes the method proposed in [RGS03] and allows us to analyze a far greater variety of actions.

2.3 Human Joint Tracking

The motivation for our approach comes from the need for detection and tracking of human body joints in higher level analysis, where action recognition has been one of the popular approaches. Many of the well received approaches [PC03, GSS04, YS05c] in this area rely on the tracking of critical human body joints. Several approaches have been proposed to solve this problem through the estimation of the pose of the human body. Not all of them generalize well to non-trivial human actions and most of them breakdown under variations in viewpoints, execution style, occlusion, etc. A recent survey by Moeslund *et al.* [MHK06] provides a detailed list of the related work and covers a variety of approaches related to human motion analysis. These approaches can be broadly categorizing into model-free and model-based approaches.

Among the model-free approaches, the bottom up detection and tracking of human body parts has always been popular [DZ05, MOB05, RMR04]. Body parts are detected using AdaBoost [MOB05], 2D shapes [RMR04], SVM classifiers [RT02], and local appearance models [DZ05]. Ramanan *et al.* [DZ05] detect the pose for a person walking or running using the scissor-leg model. In [XM05], Ren *et al.* propose pairwise constraints (aspect ratio, scale, appearance, orientation, and connectivity) between body parts to assemble detected body parts into 2D configurations. However, model-free approaches are challenged by the complexity of human body motion, appearance, and pose. Hence, model based approaches make use of prior information about the structure and kinematics of the human body. Researchers have used 2D and 3D models for this task. Our approaches falls into the first group. Depending on the challenges and requirements of different domains, researchers have proposed shape based approaches [LSS05, WY06], motion based approaches [BD02, Sid04], and a combination of both types [KKP03, WN05]. Most of these approaches are very sensitive to a change in viewpoints. To overcome this limitation, most of the methods rely on the use of a large database of shapes and motion patterns that is either used for training a model or as a look-up table. Even with the availability of a representative data set, approaches of this type are limited to simple human actions. The analysis of a complex human motion requires more sophisticated methods that rely on useful features extracted from the shape or motion of the human body [EBM03, RCK06]. However, the performance of these approaches rely on contour and motion features that are sensitive to view changes.

Some researchers [IC03, GK03] addressed the problem of direct reconstruction of both model shape and motion from the visual-hull without a prior model. Using multiple synchronized views in [IC03], Mikic *et al.* employ hierarchical approach using prior knowledge of average body part shapes and dimensions to detect and label body parts in foreground silhouettes. Once the model is initialized, an extended Kalman filter is used to estimate model parameters between frames. In [GK03], Cheung *et al.* use multiple calibrated views to correctly obtain foreground silhouettes and reconstruct a model of the kinematic structure, shape, and appearance of a person. The acquired shape and joint information is then used to track the motion of the person in other sequence of images.

Among the 3D model based approaches, both monocular and multiple view methods have been used, although monocular is usually preferred for practicality. These approaches are also useful for synthesis using the detections. Sminchisescu *et al.* [ST03] use stochastic sampling for recovering 3D pose in the case of monocular view. Multiple views have been used by Plankers *et al.* [PF02] to track the upper body using stereo pair and silhouette cues. The essence of this type of approach is to predict the pose of the model corresponding to the foreground silhouette. This is done for a number of predicted model poses until the best match is found. Obviously, the search space contains a very large number of possible model poses.

Assuming smoothness of a human motion, in [GS06], Gritai *et al.* applied an epipolar line constraint between human bodies for tracking joints. The approach cannot be used for the prediction of trajectories because of the uncertainty in computing the true landmark locations along the epipolar line. The approach presented in this paper resolves this ambiguity by using the affine constraint for the prediction of the complete action in the test video. The geometrical formulation of human action for the task of body joint detection and tracking is novel. Our approach does not require extensive training, generalizes well to complicated articulated actions, and is robust to variations in anthropometry of individuals, viewpoints, execution rate, and the style of action. In addition to this, we can also handle occlusion of limbs as captured by the model.

CHAPTER 3
INVARIANCE IN ACTION RECOGNITION

An action can be represented by series of stick figures. It is known that this representation contains sufficient information for humans to recognize an action. Based on a single trajectory action representation, an action dissimilarity measure is proposed. The measure is derived from the epipolar geometry between two views. Observing that anthropometric measurements do not vary arbitrary, the *postural* and *action* constraints are explicitly stated. Based on these anthropometric constraints, the proposed dissimilarity measure is extended to the matching of multiple trajectories for human actions. Since the temporal execution rate of human action varies, Dynamic Time Warping was utilized to match actions performed by different humans. The proposed method was validated on several image sequences containing human actions.

3.1 Action Representation

The model of a moving body as a point is ubiquitous in the physical sciences community. In our work, the input is the 2D motion of a set of 13 anatomical landmarks, $\mathcal{L} = \{1, 2, \cdots 13\}$, as viewed from a camera (see Figure 3.1). In [Joh73], Johansson demonstrated that a simple point-based model of the human body contained sufficient information for the recognition of actions. Relying on this result, we represent the current pose and posture of an actor in terms of a set of points in 3D-space $\hat{\mathbf{X}} = \{\mathbf{X}_1, \mathbf{X}_2 \dots \mathbf{X}_n\}$, where $\mathbf{X}_i = (X_i, Y_i, Z_i, \Lambda)^\top$ are homogenous coordinates. A *posture* is a stance that an actor has at a certain time instant, not to be confused with the actor's *pose*, which refers to position and orientation (in a rigid sense). Each point represents the spatial coordinate of an anatomical landmark on the human body as shown in Figure 3.1. For the k^{th} camera, the imaged pose and posture are represented by $\mathbf{U}^k = \{\mathbf{u}_1^k, \mathbf{u}_2^k \dots \mathbf{u}_n^k\}$, where $\mathbf{u}_i^k = (u_i, v_i, \lambda)^\top$. $\hat{\mathbf{X}}$ and $\mathbf{U}^k$ are related by a 4×3 projection matrix $\mathtt{C}^k$, i.e. $\mathbf{U}^k = \mathtt{C}^k \hat{\mathbf{X}}$. As will be seen, eight points on human body are required in each frame of video, and at least one of them must correspond to the body part directly involving in action. We refer to each entity involved in an *action* as an *actor*. An *action element* is the portion of an action that is performed in the interval between two frames. Each action is represented as the set of action elements. For a comparison of other representations to this one the reader is referred to [Gav99].

3.2 Viewpoint Transformations

Figure 3.2 shows the same action ('picking up a book') from four different points of view. Although the same action is being performed, the distribution of points on the image differs significantly. As has been observed previously for object recognition, it is usually unreasonable to place restrictions on the possible viewpoint of the camera, and action recognition algorithms should therefore demonstrate *invariance* to changes in viewpoint. Invariants are properties of geometric configurations that are unaffected under a certain class of transformations. It is known that absolute

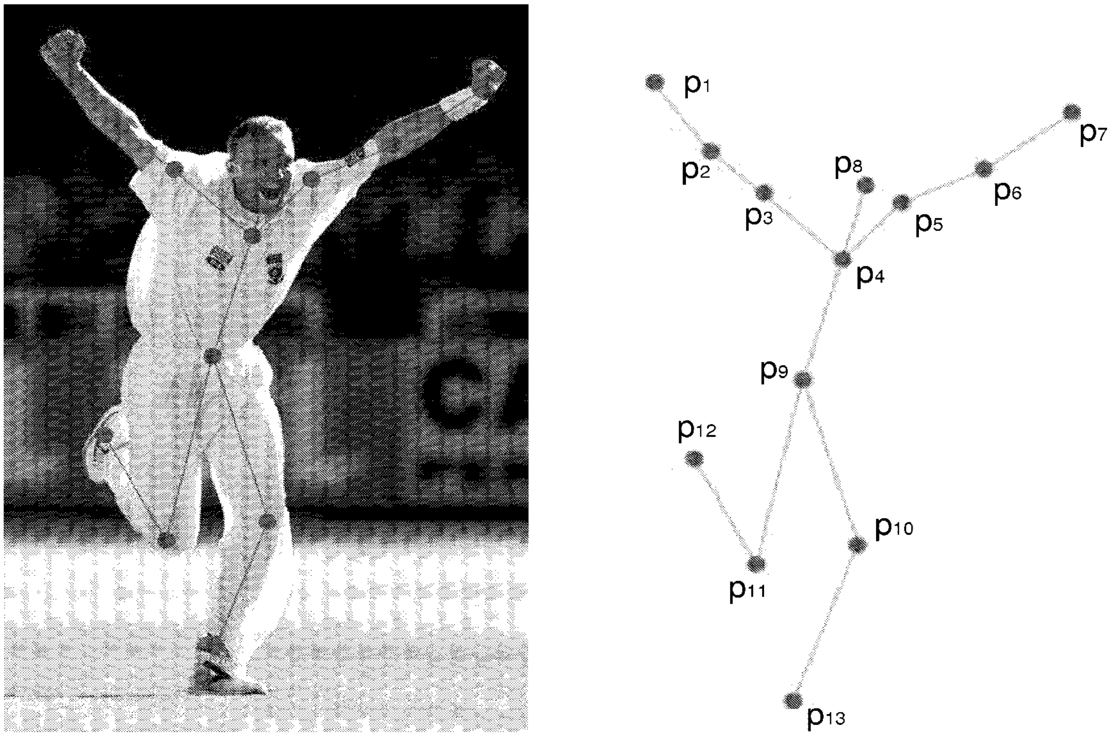


Figure 3.1: Point-based representation. Johansson's experiments in demonstrate that point-based representations contains sufficient information for action recognition.

invariants do not exist for general 3D point sets [BWR92]. However, there are useful properties that are not strictly invariant, but remain stable over most transformations.

We now describe a measure to match actions that is based on one such property. Assuming two frames are temporally aligned (until Section 3.4), the labels associated with each anatomical landmark provide point-to-point correspondence between the two postures. The constraint we use is that if the two imaged point sets match, they are projections of the same structure in 3D. In [RS01], a rank constraint based dissimilarity measure was described that was stable to camera viewpoint changes. The main drawback of this dissimilarity measure was the assumption of affine cameras. To remove this assumption, instead of using this factorization based rank constraint, we

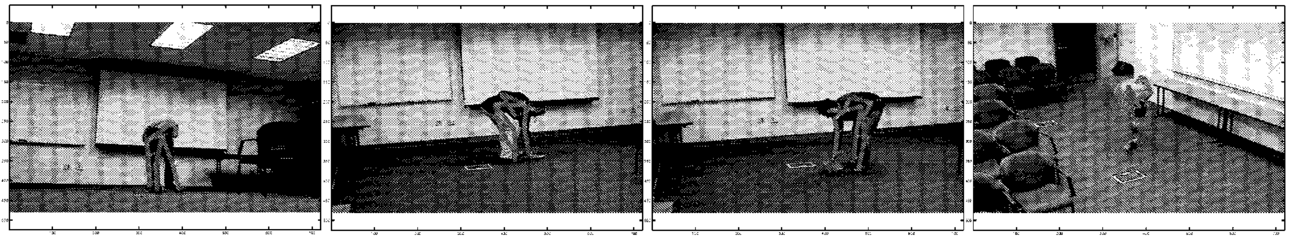


Figure 3.2: Frames corresponding to 'picking up' in four sequences. The left-most frame corresponds to the model sequence, and the rest correspond to the test sequences. In each sequence, the actors are in markedly different orientations with respect to the camera, but in the same posture.

use a constraint derived from epipolar geometry. For the projective camera model, the fundamental matrix (a 3 × 3 matrix of rank 2), $\mathbf{F}$, is defined between corresponding points by

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix}^T \mathbf{F} \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = 0, \tag{3.1}$$

for the pair of matching points $(u, v) \leftrightarrow (u', v')$ in trajectories, observed from two different viewpoints. Clearly, given a fundamental matrix, we can use Equation 3.1 to measure the dissimilarity between two trajectories so that the squared residual for all points is minimized. By rearranging Equation 3.1, a dissimilarity measure can also be defined directly from the trajectory values themselves (without explicitly computing $\mathbf{F}$). Given at least 8 point matches, we have,

$$\mathcal{A}\mathrm{f} = \begin{bmatrix} u'_1 u_1 & u'_1 v_1 & u'_1 & v'_1 u_1 & v'_1 v_1 & u'_1 & u_1 & v_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u'_t u_t & u'_t v_t & u'_t & v'_t u_t & v'_t v_t & u'_t & u_t & v_t & 1 \end{bmatrix} \mathrm{f} = 0, \tag{3.2}$$

where $\mathbf{f} = \begin{bmatrix} f_{11} & f_{12} & f_{13} & f_{21} & f_{22} & f_{23} & f_{31} & f_{32} & f_{33} \end{bmatrix}^T$ is the fundamental matrix vectorized in row-major order. We refer to $\mathcal{A}$ as an observation matrix, which is constructed using only the coordinates of points of corresponding 2D trajectories. Since Equation 3.2 is homogenous, for a non-trivial solution of $\mathbf{f}$ to exist, matrix $\mathcal{A}$ must have a rank of at most eight, and this can be exploited to measure dissimilarity. Of course, due to the noise or the matching error, the rank of matrix $\mathcal{A}$ may not be exactly eight. The number, which we will call *kappa*, inverse to the condition number of $\mathcal{A}$ (the ratio of the smallest singular value, σ_9, to the largest singular value, σ_1) of $\mathcal{A}$ provides the algebraic error of corresponding points in matrix $\mathcal{A}$. This ratio can be used to measure the match of two trajectories,

$$\kappa = \frac{\sigma_9}{\sigma_1}. \tag{3.3}$$

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

3.3 Anthropometric Transformations

Both body size and proportion vary greatly between different races and age groups and between both sexes. However, while human dimensional variability is substantial, several anthropometric studies (see [EKC82], [Bri95], [BPW93]) empirically demonstrate that it is not *arbitrary*. These studies have tabulated various percentiles of the dimensions of several human anatomical landmarks. In this paper, we conjecture that for a significant portion of the human population the proportion between human body parts coupled with a rigid transformation in 3D space can be captured by a projective transformation of $\mathbb{P}^3$.

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

$$\mathbf{X}_i = \mathcal{M}\mathbf{Y}_i \tag{3.4}$$

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

This was empirically supported using the data in [Bri82] (see Table $5 - 1$ and $5 - 2$ which record the body dimensions of male and female workers between the ages of 18 and 45). Between the dimensions of the '5^{th} percentile woman' and the '95^{th} percentile man', where a mean error of 227.37 mm was found before transformation, a mean error of 23.87 mm was found after applying an appropriate transformation. Using this property, geometric constraints can be used between the imaged points, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, of the two actors. The transformation $\mathcal{M}$ simultaneously captures the different pose of each actor (with respect to a world coordinate frame) as well as the difference in size/proportions of the two actors.

3.3.1 Postural Constraint

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

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

It is known (pg. 247 Section 9.2, [HZ00]) that for uncalibrated cameras, the ambiguity of structure is expressed by such an arbitrary non-singular projective matrix. If two actors are in the same posture, the only difference between their point-sets is a projective relationship (Conjecture 1). Thus, if an invertible matrix $\mathcal{P}$ exists between $\mathbf{X}$ and $\mathbf{Y}$, i.e. $\mathbf{X} = \mathcal{P}\mathbf{Y}$, a *fundamental* matrix is uniquely determined by $\mathbf{x}^\top \mathcal{F} \mathbf{y} = 0$ (Theorem 9.1 [HZ00]).[1] It is important to note that the matrix $\mathcal{F}$ does not capture only the relative positions of the cameras as does the fundamental matrix $\mathbf{F}$, but instead the relative poses of the actors and the relative anthropometric transformation between the actors.

Since the labels of each point are assumed known, *semantic* correspondences (i.e. the left shoulder of A_1 corresponds to the left shoulder of A_2) between the set of points are also known. Proposition 1 states that the matrix computed using these semantic correspondences between actors inherently captures the difference in anthropometric dimensions and the difference in pose. This point is illustrated in Figure 3.3. The matrix $\mathcal{F}$, computed between the actors, captured an

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

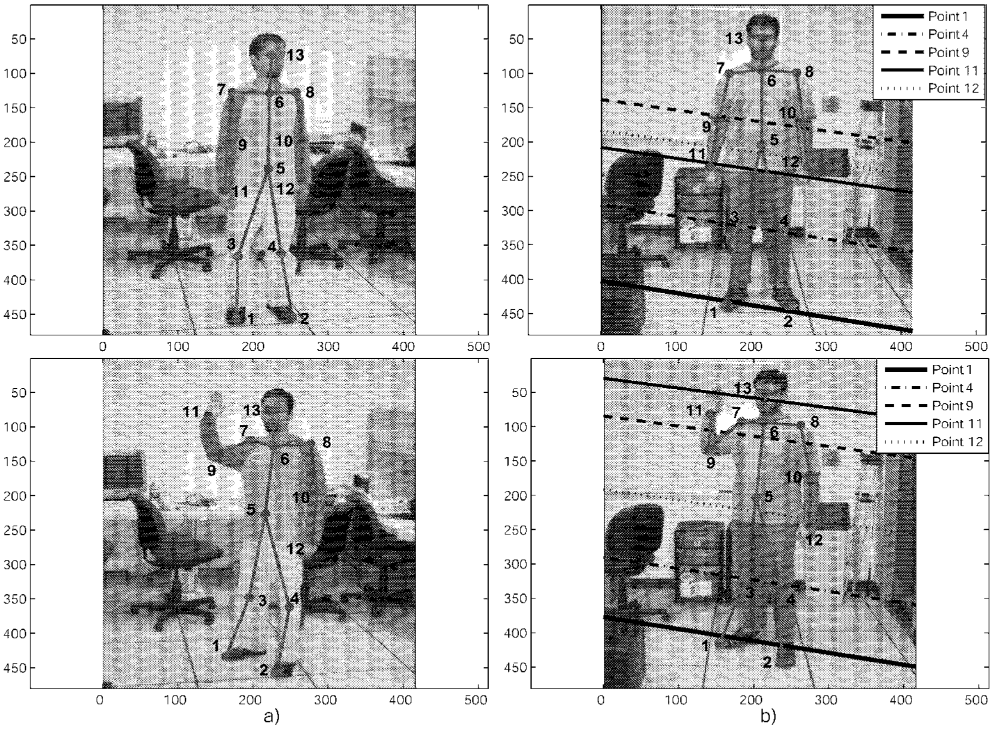


Figure 3.3: The matrix $\mathcal{F}$ can capture the relationship between body joints of two different actors of different height, weight, etc. but in the same posture. It captures the variability in proportion as well as the change in viewpoint. (a) Actor 1 in two frames of the model video. (b) Actor 2 in the corresponding frames of the test video. The joint correspondences in first frames of model and test video were used to compute the matrix $\mathcal{F}$. The image on right in (b) shows epipolar lines corresponding to joints in the image on right in (a). It is clear that the joints in the test video lie close to the corresponding epipolar lines.

anatomical relationship between the actors as well as the different views of the actors. The result is that the dissimilarity measure, described in Section 3.2, remains stable despite changes in anthropometry of the actors. Since the anthropometric proportions of actor can be expected to remain the same over short periods of time this fact can be used to provide an even stronger constraint which we now describe.

3.3.2 Action Constraint

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

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

Based on Conjecture 1, we can say that $\mathcal{M}$ remains the same between time t and $t+1$. In other words, $\mathcal{M}$ determines $\mathbf{Y}$ with respect to $\mathbf{X}$ and does not depend on the motion of $\mathbf{X}$. Since $\mathcal{M}$ is the same then the matrices, $\mathcal{F}_t$ and $\mathcal{F}_{t+1}$, corresponding to $(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t)$ and $(\hat{\mathbf{x}}_{t+1}, \hat{\mathbf{y}}_{t+1})$ are the same (p.235 Result 8.8, [HZ00]).

This means that if both individuals perform the same action-element between frame f_t and frame f_{t+1}, the transformation that captured the difference in pose and dimension between the two actors remains the *same*. As a direct consequence, the subspace spanned by the measurement matrix A also remains the same and this suggests that if a measurement matrix were constructed using *all* the corresponding points over the entire action $\hat{A} = [A_1, A_2, \ldots A_k]$, $\kappa_{\hat{A}}$ can be used as a global measure of *action* dissimilarity. The second row of Figure 3.3 illustrates this. Both actors of clearly different anatomical proportion perform the same action element (they moved their right foot back and raised their right hand). The matrix $\mathcal{F}$ computed between the actor in their original postures was used to compute epipolar lines after the execution of the action element. Clearly, to the extent that the same action element was performed, the geometric relationship is preserved.

Thus, instead of considering the action as the successive motion of 13 points over n frames, each action is considered to be a cloud of $13n$ points, each point having a unique spatio-temporal index (see Figure 3.4). However, the analysis thus far has assumed that temporal transformations had been accounted for. In practice, temporal transformations, small or large, always exist. We now describe how to compensate for these transformations during action analysis.

3.4 Temporal Transformations

While invariance to change in viewpoint is required in action analysis due to the imaging process, invariance to temporal transformations is needed due to the nominal uniqueness of each actor's execution of an action. Here, we propose a matching algorithm that is stable under nonlinear type of transformations. We describe a general approach, where the temporal transformation may be highly nonlinear, using DTW to compensate for temporal transformations. In this case, there is no clearly defined class of temporal transformations, except that temporal order must be preserved during the transformation.

3.4.1 Non-linear Transformation

Dynamic Time Warping is a widely used method for warping two temporal signals [SC78]. It uses an optimum time expansion/compression function to perform a non-linear time alignment. The applications include speech recognition, gesture recognition [DEP95], signature verification

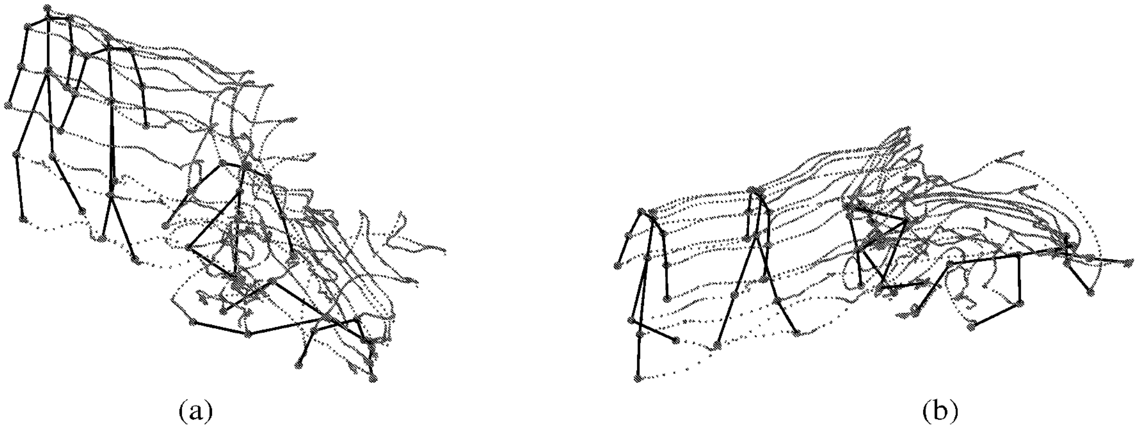


Figure 3.4: Cloud of points for the action 'getting up' from two views by actors of different anthropometric proportions. Instead of considering the action as the successive motion of 13 points over n frames, each action is considered to be a cloud of $13n$ points, each point having a unique spatio-temporal index. (a) View 1 (b) View 2.

and for video alignment [RGS03]. DTW is particularly suited to action recognition, since it is expected that different actors may perform some portions of an action at different rates, relatively. The use of DTW is not trivial in this case since both the local (postural) constraint and the global (action) constraint need to be incorporated in the computation of the dissimilarity measure. Applying a temporal window (k frames before and after the current one) for computation of dissimilarity measure between two agents provided a marked improvement.

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

$$\mathbf{E}(i, j) = \mathbf{dist}(i, j) + \mathbf{e}, \tag{3.5}$$

where

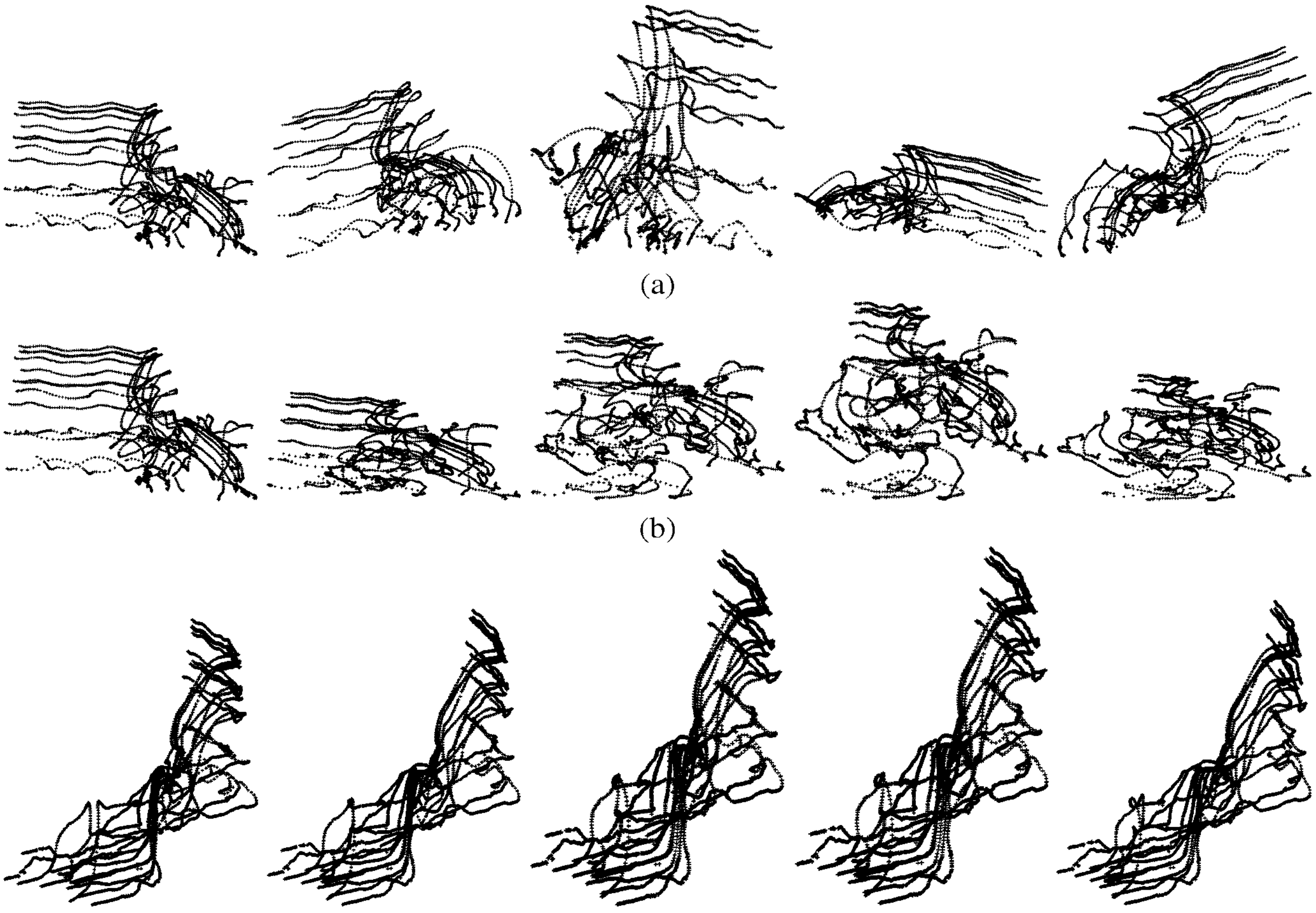
$$\mathbf{e} = \min\left\{\mathbf{E}(i-1,j), \mathbf{E}(i-1,j-1), \mathbf{E}(i,j-1)\right\}.$$

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

Similar to [RGS03], in our framework, I and J are trajectories representing similar or different actions observed from distinct viewpoints. By introducing $\kappa(i,j)$ as the $\mathbf{dist}(i,j)$, the standard DTW becomes appropriate for action recognition and robust to view, anthropometric, and temporal transformations.

3.5 Results

To validate the proposed work, we performed experiments on motion capture data and several realistic challenging scenarios. All data used during recognition was in the form of image measurements from uncalibrated cameras. The first set of experiments was performed on Motion Capture data. The second set of experiments involved action detection in a long sequence, the third set involved synchronizing videos to match actions, and the forth set applied the proposed


(a)

(b)

(c)

Figure 3.5: The trajectories of anatomical landmarks of the 'getting up' action under different types of transformation. The first row (a) presents the trajectories under different viewpoint transformations, the second row (b) under anthropometric transformations, and the third row (c) demonstrates the same trajectories obtained with different camera velocities along the $X-$direction.

3.5.1 Viewpoint

In this set of experiments we tested the performance of the system with respect to changes in viewpoint. We demonstrated that the dissimilarity measure allows sufficient discrimination between matches and mismatches, despite different viewpoints. The first row of the Figure 3.5

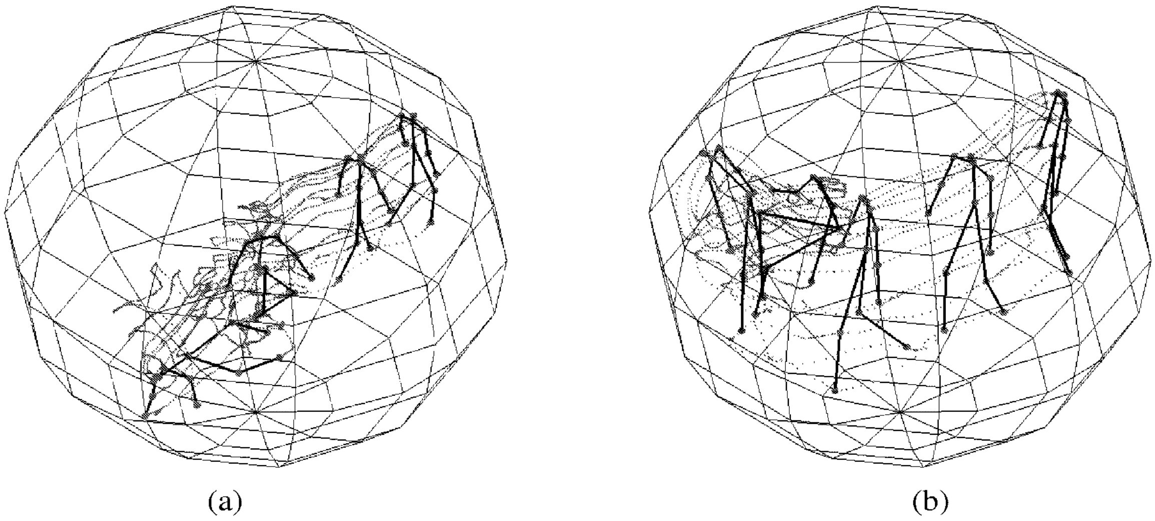


Figure 3.6: Viewing spheres. (a) The action 'getting up' is viewed at regular intervals on a sphere around the action. (b) The action 'Sit Down' is viewed at regular intervals on a sphere around the action.

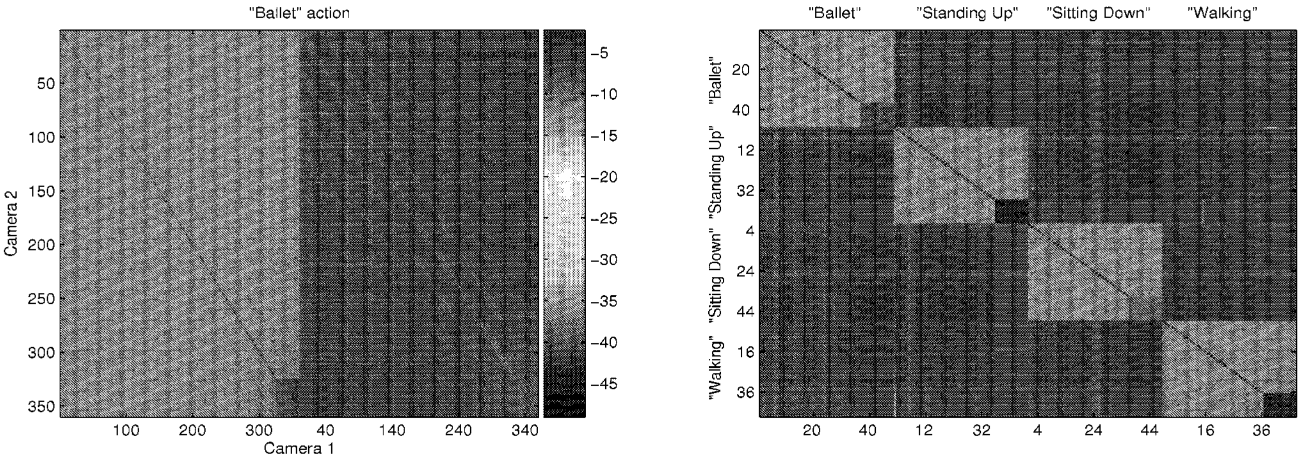
shows the input point cloud, representing the 'getting up' action, under different view projective transformations. The experimental performance is also tested with respect to increasing noise in the measurements.

Motion capture data was used to provide 3D data. Since the 3D coordinates of the points were known, 2D image coordinates were obtained by generating projection matrices around a viewing sphere as shown in Figure 3.6. The action was observed from 360 different locations in the upper hemisphere, which means the elevation and azimuth were changed from 0 to 90 and from 0 to 350 degrees respectively with a ten degree incremental step. Thus, a pair of angles, elevation and azimuth, correspond to any of 360 possible camera locations. The elevation and azimuth corresponding to some camera location n, where $n = 1, \ldots, 360$, can be calculated as

$\mathrm{floor}((n-1)/36) \times 10$ and $\mathrm{mod}((n-1)/36) \times 10$ respectively, e.g., if n=239, then the elevation and azimuth are 60 and 220 degrees respectively.

We experimented with no-noise, 'exact', data. To demonstrate the robustness to changes in viewpoint, we recorded the log of the ratio of the first smallest singular value to the largest singular value and the log of the ratio of the second smallest singular value to the largest singular value of $\mathcal{A}$ in Figure 3.7 (a). This figure shows that regardless of view angles the dissimilarity measure (left half of the matrix or first 360 values on the horizontal axis) is very close to zero and significantly lower than the ratio of the second smallest singular value to the largest singular value (right half of the matrix). From the illustration, one can notice that the diagonal elements are especially low. The diagonal entities correspond to the case when both camera views are exact the same. Within this matrix, there are blocks of low values (indices of both axes are between 325 and 360). These values correspond to the case when elevations of both cameras are facing to the ground, 90 degrees, and is a degenerate case. From our experiment, in the degenerate case, the values of the two ratios are approximately 4.3×10^{-22} and 1.3×10^{-20}, while in all other matches the mean of the dissimilarity measure is 1.4×10^{-16} and the mean of the other ratio is 5.4×10^{-4}.

Figure 3.7 (b) shows a confusion matrix using $\log \kappa$ in a second series of experiments, where different actions were compared to each other. Four actions ('ballet', 'standing up', 'sitting down' and 'walking') were rendered from 360 different viewpoints and the block diagonal structure of the confusion matrix shows the discrimination achieved using the proposed measure. It is important to note that even in the degenerate case mentioned above, *kappa* provides ample discrimination between different actions.



(a)

(b)

Figure 3.7: Four different actions were compared to itself. Pattern and test actions were observed from any angle of the upper hemisphere. The left-most figure shows a significant drop between two ratios σ_9/σ_1(blue) and σ_8/σ_1(red), thus σ_9/σ_1 can be considered as a dissimilarity measure. When both cameras have the elevation angle of 90 degrees, which corresponds to the upper point of hemisphere (325 and 350 indices), both ratios are very low. Since at the upper point of hemisphere camera centers coincide, it becomes a degenerate case. The right-most figure shows the change of σ_9/σ_1, when under different view-projective transformations, four different actions were compared to each other. The low diagonal values of the proposed dissimilarity measure demonstrate the correct discrimination among actions.

3.5.2 Action Recognition

In this experiment, actors performed a sequence of actions: 'walking', 'picking up an object', 'lifting the object', and 'walking' away. Videos were taken of two different actors each from two different views, and the action of picking up an object was detected in each video by matching a shorter sequence containing only 'picking up an object'. Figure 3.8 shows the corresponding frames in the four videos. The sensitivity of recognition was also tested in a sequence containing four individuals walking. A test pattern of a single cycle of the distinctive 'Egyptian' gait was compared to each actor's motion and the variation of the *kappa* over time for each of the four

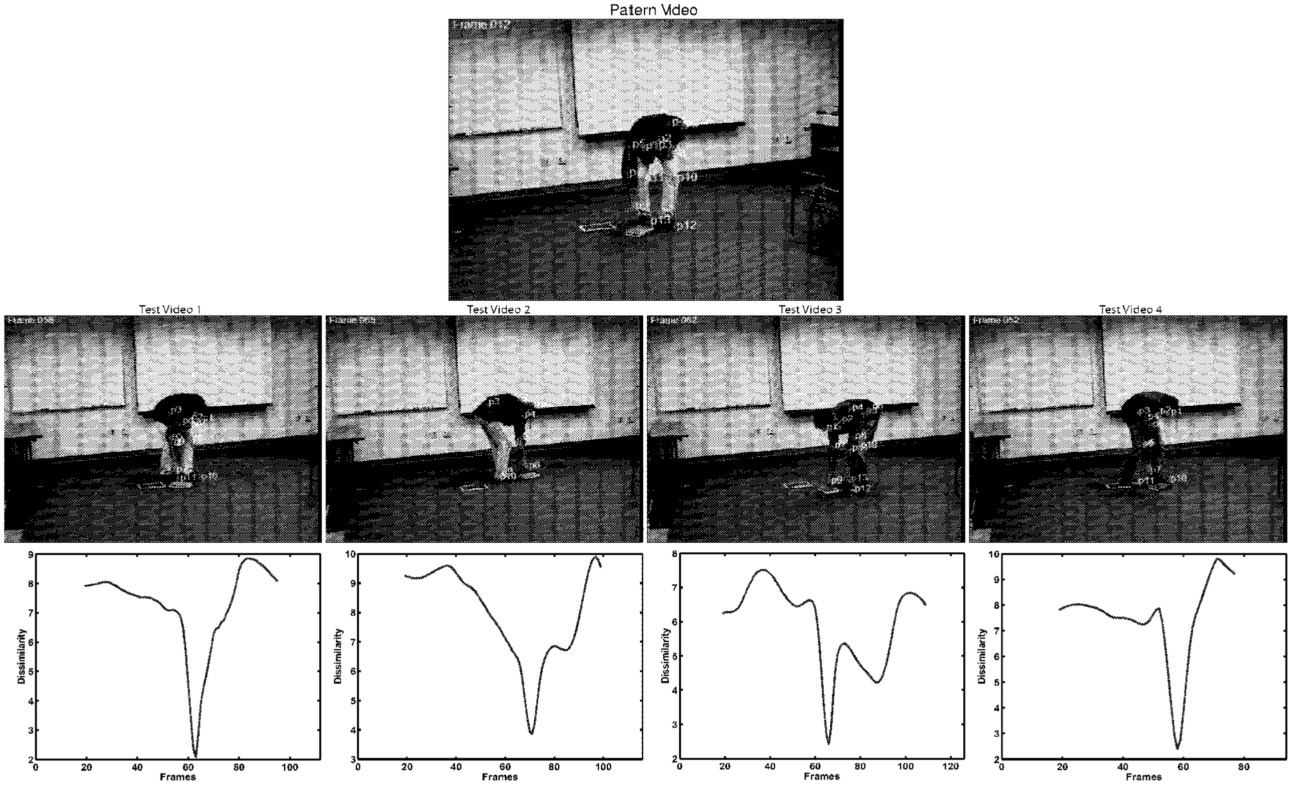


Figure 3.8: Action detection from multiple views. Action was performed by two actors, and captured from several distinct view points. All corresponded postures were detected successfully.

actors is shown in Figure 3.9 (the odd-one-out is the third actor from the left). There are two interesting points that can be observed in this figure. Firstly, since the posture involved in the 'Egyptian' gait is relatively distinct from the usual human gait, the dissimilarity measure for the third actor is consistently larger and distinct from the other actors. Secondly, the sinusoidal nature of the plot clearly shows the periodicity that is associated with walking.

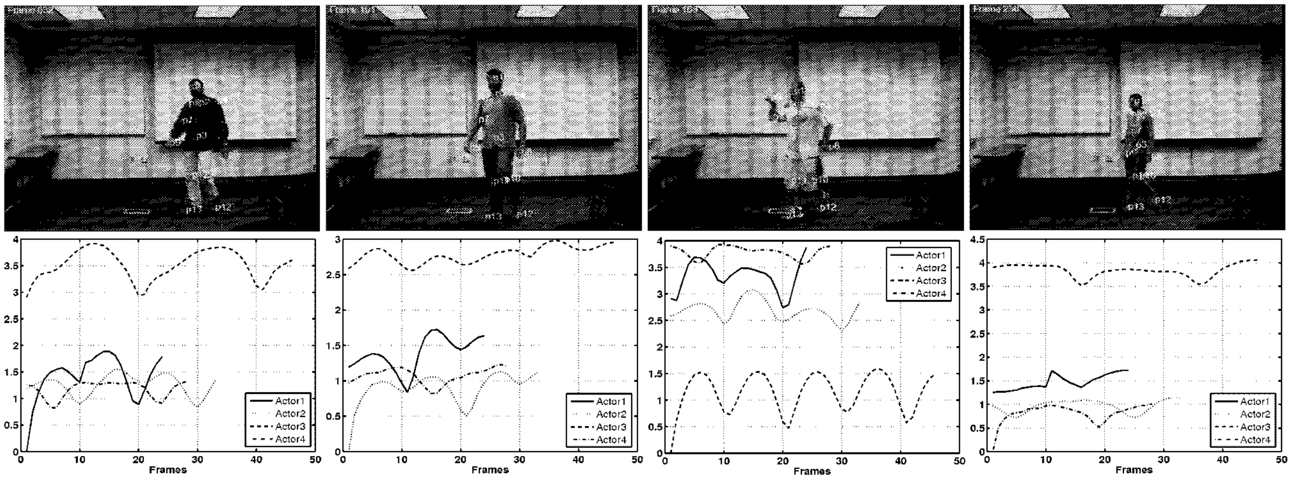


Figure 3.9: Finding the Odd Man Out. Actor three, the third figure from the left, corresponds to the actor performing the 'Egyptian' gait.

3.5.3 Action Synchronization

Three actors jumped asynchronously in the field of view of a stationary camera. The objective in this experiment was to align the actor's jumps and twists so that a new synchronized sequence could be rendered. The temporal transformation between actors was highly nonlinear, and DTW with a 10-frame window around the current frame was used. Using the proposed approach , the accurate synchronization was achieved and Figure 3.10 shows the result of the synchronization with respect to the left-most actor. The top row shows the original sequence and the bottom row shows the rendered sequence. The two other actors were synchronized with respect to the left-most actor individually.

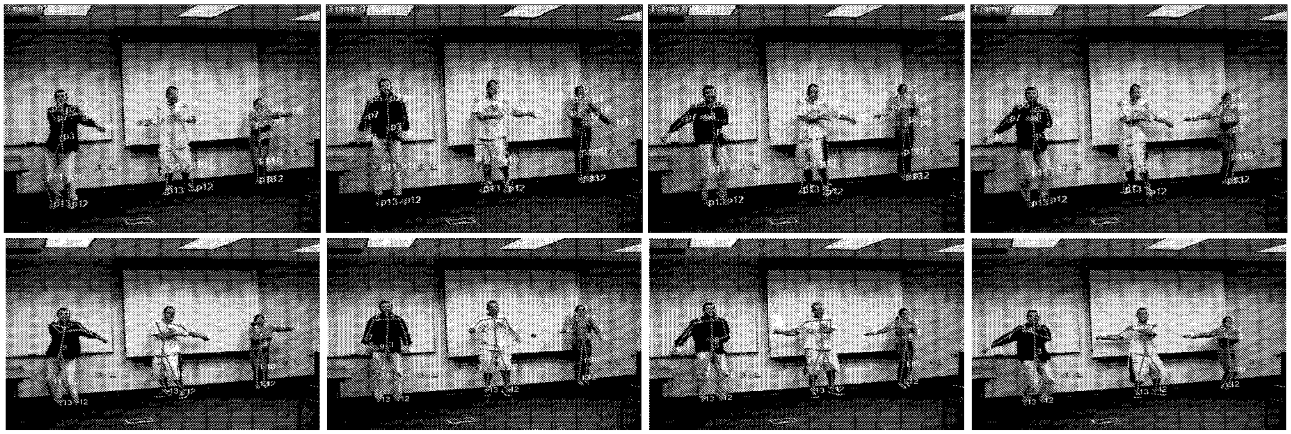
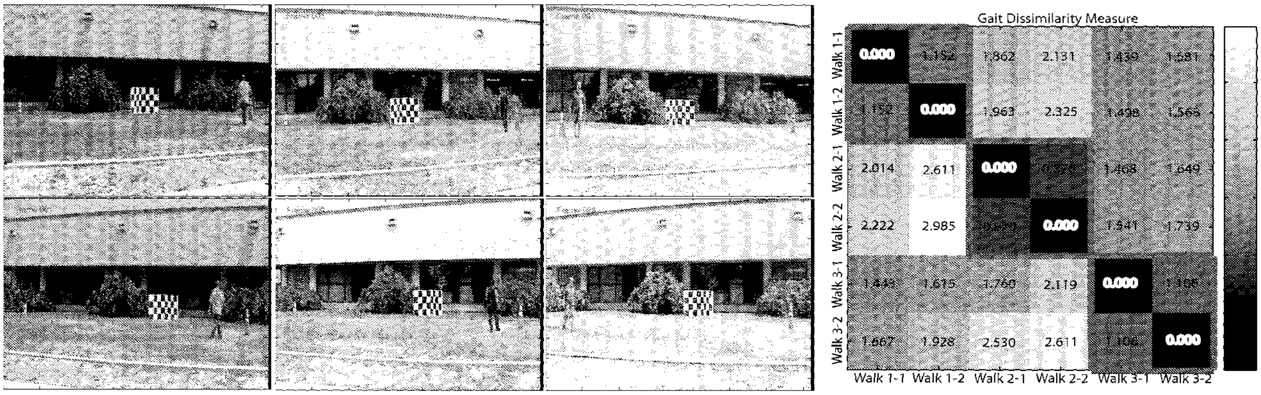


Figure 3.10: Following the leader (the left-most actor). The top row shows four frames, 22, 25, 27, and 29 before synchronization. Notice the difference in postures of each actor within each single view. The bottom row shows corresponding frames (to the top row) from the rendered sequence after synchronization.

3.5.4 Gait Analysis

Three walking actors were captured from two different viewpoints using two cameras, and on average, each video was more than 200 frames in length. Six feature points, hands, knees, and feet were tracked. A short fragment (40 frames) was extracted from each video. The goal of experiment was to determine if the extracted fragment could be found in the video by computing the κ as the best dissimilarity measure. The table of Figure 3.11 shows the confusion matrix of each gait in each view. In the table, the first and second columns correspond to the first actor in the first and second view, respectively, and so on. As expected, the distance between the gait of an actor in first view and in the second view is always lower than the gait of other actors in any view.



	Walk 1-1	Walk 1-2	Walk 2-1	Walk 2-2	Walk 3-1	Walk 3-2
Walk 1-1	0.000	1.152	1.862	2.131	1.439	1.581
Walk 1-2	1.152	0.000	1.963	2.325	1.498	1.566
Walk 2-1	2.014	2.611	0.000	[illegible]	1.468	1.649
Walk 2-2	2.222	2.985	[illegible]	0.000	1.541	1.739
Walk 3-1	1.443	1.615	1.760	2.119	0.000	1.106
Walk 3-2	1.667	1.928	2.530	2.611	1.106	0.000

Figure 3.11: Confusion Matrix for Gait Analysis. Three walking actors were captured from two different view points using two cameras. The table shows the confusion matrix of each gait in each view. In the table the first and second columns correspond to the first actor in the first and second view respectively, and so on. The notation 1-1 refers to 'Actor 1, View 1' etc. Note that lower values correspond to the same actor's gaits in different views (1-1 matches best with 1-2, 2-1 with 2-2, 3-1 with 3-2)

3.6 Summary

We propose a method for matching human actions using semantic correspondence between human bodies. The correspondences are used to provide geometric constraints between multiple anatomical landmarks (e.g. hands, shoulders and feet) to match actions performed from different viewpoints and in different environments. The fact that the human body has certain anthropometric proportion allows innovative use of the machinery of epipolar geometry to provide constraints to accurately analyze actions performed by different people leading to some interesting results. To our knowledge, this is the first work that expressly addresses the variability of human proportions. Temporally invariant matching is performed, using non-linear time warping, to ensure that similar actions performed at different rates are accurately matched as well. Thus, the proposed algorithm guarantees that both temporal and view invariance is maintained in matching. We demonstrate the versatility of our algorithm in a number of challenging sequences and applications.

CHAPTER 4
SPACETIME PROJECTION FOR UNIFORMLY MOVING CAMERAS

In the previous chapter, we discussed an approach for human action recognition in videos acquired by a fixed camera. In the real world, a camera may mounted on a vehicle such as aircraft, train, car, or on a robot. Therefore, it is important to be able to recognize human actions in videos acquired by a moving camera. In this chapter, we consider cameras moving at constant velocity in 3D space. We introduce the concept of spacetime projection and derive the fundamental matrix between two such cameras, from which six different fundamental matrices can be directly recovered. In a similar manner, to which the standard fundamental matrix was applied to action understanding, the fundamental matrix between moving cameras is used to recognize actions. Qualitative and quantitative experiments demonstrate the performance of the proposed measure.

4.1 Spacetime Projection Model

We define a *world point* as $\mathbf{X} = [T\ X\ Y\ Z]^T \in \mathcal{R}^4$, on a world coordinate $\mathbf{U} = [T\ \lambda X\ \lambda Y\ \lambda Z\ \lambda]^T$ and a *video point* as $\mathbf{X} = [t\ u\ v] \in \mathcal{R}^3$ on a video coordinate system $\mathbf{u} = [t\ wu\ wv\ w]^T \in \mathcal{R}^4$. Assuming square pixels and that the world and camera coordinate systems are aligned, the mapping describing central projection for the spatial coordinates and orthographic projection for the temporal coordinate are,

$$(T, X, Y, Z)^T \mapsto (\alpha_t T, fX/Z + p_u, fY/Z + p_v)^T \tag{4.1}$$

where f is the focal length of the camera and α_t is the reciprocal of the frame-rate of the camera (causing an effect akin to time dilation) and (p_u, p_v) are the coordinates of the principal point. This can be expressed in matrix form as

$$\begin{bmatrix} T \\ X \\ Y \\ Z \end{bmatrix} \mapsto \begin{bmatrix} t \\ wu \\ wv \\ w \end{bmatrix} = \begin{bmatrix} \alpha_t & & & \\ & f & & p_u \\ & & f & p_v \\ & & & 1 \end{bmatrix} \begin{bmatrix} T \\ X \\ Y \\ Z \end{bmatrix}, \tag{4.2}$$

or more concisely $\mathbf{u} = \mathbf{KX}$, where $\mathbf{K}$ is the calibration matrix. Generally, the spatial world and camera coordinate systems are related by rotation and translation and the temporal coordinates by a translation (e.g. the time index when camera begins recording). These transformations can be captured by a 4×4 orthogonal matrix $\mathbf{Q}$ and a 4×5 displacement matrix $\mathbf{D}$, where

$$\mathbf{Q} = \begin{bmatrix} 1 & \\ & \mathbf{R} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} 1 & & & & 0 \\ & 1 & & & -D_x \\ & & 1 & & -D_y \\ & & & 1 & -D_z \end{bmatrix}, \tag{4.3}$$

where $\mathbf{R}$ is a 3×3 rotation matrix capturing the orientation of the camera coordinate system and $\mathbf{C} = [0, D_x, D_y, D_z]^T$ is the position of the camera center. The 4×5 projection matrix relates the world and video coordinate systems, $\mathbf{u} = \mathbf{PU}$. This projection matrix can be decomposed as

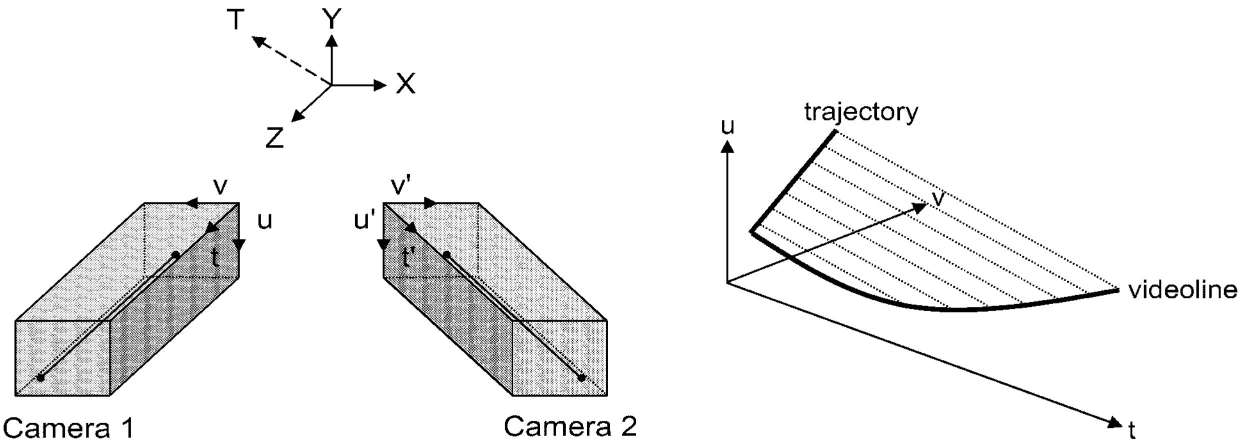


Figure 4.1: Galilean Cameras.(a) Projection onto the video hyperplane (b) The videoline of a point charts out a hyperbolic function in spacetime.

$\mathbf{P} = \mathbf{KQD} = \mathbf{KQ}[\mathbf{I}| - \mathbf{C}]$ or simply $\mathbf{P} = \mathbf{K}[\mathbf{Q}| - \mathbf{QC}]$. If the cameras are moving at constant velocity according to $\Delta\mathbf{C} = [0, \Delta D_x, \Delta D_y, \Delta D_z]^T$, we have the following series

$$\begin{aligned} \mathbf{u}(0) &= \mathbf{KR}[\mathbf{I}| - \mathbf{C}]\mathbf{U} \\ \mathbf{u}(1) &= \mathbf{KR}[\mathbf{I}| - (\mathbf{C} + \Delta\mathbf{C})]\mathbf{U} \\ &\vdots \\ \mathbf{u}(T) &= \mathbf{KR}[\mathbf{I}| - (\mathbf{C} + T\Delta\mathbf{C})]\mathbf{U}. \end{aligned} \tag{4.4}$$

Now, including the temporal dimension into the object vector we can rewrite this simply as,

$$\mathbf{u} = \mathbf{KQ}[\mathbf{G}| - \mathbf{C}]\mathbf{U}, \tag{4.5}$$

where

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\Delta D_x & 1 & 0 & 0 \\ -\Delta D_y & 0 & 1 & 0 \\ -\Delta D_z & 0 & 0 & 1 \end{bmatrix}$$

is a Galilean transformation. We refer to $\mathcal{M} = \mathbf{KQ}[\hat{\mathbf{D}}| - \mathbf{C}]$ as the *spacetime projection matrix*. As with the spatial projection matrix, the null vector of $\mathcal{M}$ corresponds to the spacetime location of the camera center in the world at $t = 0$. In addition, $m_{12} = m_{13} = m_{14} = 0$, where m_{ij} is the element in the *i*th row and *j*th column of $\mathcal{M}$. The video taken by a Galilean camera can therefore be properly considered a three-dimensional image projected from a four-dimensional world.

Analogous to worldlines in spacetime geometry [Cal00], we refer to the curve charted out by successive events from a point in the world as *videolines*. It was shown by Bolles *et al.* in [RM87] that these curves are described hyperbolic functions on EPIs, but in the video hyperplane (assuming that the world reference frame is aligned with the camera reference frame) it follows the parametric form,

$$u(T) = p_u + \frac{-f\Delta D_x T + fX}{-\Delta D_z T + Z}, \tag{4.6}$$

$$v(T) = p_v + \frac{-f\Delta D_y T + fY}{-\Delta D_z T + Z}, \tag{4.7}$$

$$t(T) = \alpha_t T. \tag{4.8}$$

It should be noted then that straight lines in the spacetime world are not mapped to straight lines in the video hyperplane, except when the principal axis is orthogonal to the velocity vector (in which case $D_z = 0$ and Z is constant and as a result Equations 4.6 and 4.7 are linear). Thus, spatial invariants such as the cross-ratio are not preserved in spacetime. In the next sections, we study the relationship between pairs of Galilean cameras in planar and general scenes.

4.2 Planar Geometry

In this section, we describe a transformation analogous to the planar homography relating two images of a world plane. By choosing two orthogonal basis vectors that span the scene plane as the X and Y axes of the world coordinate system and ignoring the perpendicular Z coordinate (since all Z values will equal zero) we have,

$$\mathbf{u} = \begin{bmatrix} t \\ wu \\ wv \\ w \end{bmatrix} = \mathcal{M}_{\hat{4}} \begin{bmatrix} T \\ X \\ Y \\ Z \end{bmatrix}, \mathbf{u}' = \begin{bmatrix} t' \\ w'u' \\ w'v' \\ w' \end{bmatrix} = \mathcal{M}'_{\hat{4}} \begin{bmatrix} T \\ X \\ Y \\ Z \end{bmatrix}, \tag{4.9}$$

where $\mathcal{M}_{\hat{4}}$ and $\mathcal{M}'_{\hat{4}}$ are nonsingular 4×4 matrices, constructed by removing the fourth column from $\mathcal{M}$ and $\mathcal{M}'$, respectively. Therefore, there exists a transformation relating $\mathbf{u}$ and $\mathbf{u}'$, i.e., $\hat{\mathcal{H}} = \mathcal{M}'_{\hat{4}}\mathcal{M}_{\hat{4}}^{-1}$, where $\mathbf{u}' = \hat{\mathcal{H}}\mathbf{u}$. Additionally, considering time independently we see that,

$$t = m_{11}T + m14; t' = m'_{11}T + m'_{14}; \frac{t - m_{14}}{m_{11}} = \frac{t' - m'_{14}}{m'_{11}} = T,$$

from which we get the mapping $t' = h_{11}t + h_{14}$, or in other words, $h_{12} = h_{13} = 0$. As a result, we get the following functions to determine t', u', and v'.

$$t' = h_{11}t + h_{14}; u' = \frac{h_{21}t + h_{22}u + h_{23}v + h_{24}}{h_{41}t + h_{42}x + h_{43}y + h_{44}}; v' = \frac{h_{31}t + h_{32}u + h_{33}v + h_{34}}{h_{41}t + h_{42}x + h_{43}y + h_{44}}.$$

Thus, a nonsingular 4×4 matrix, $\mathcal{H}$, relates the spacetime coordinates of two videos captured by Galilean cameras observing a planar scene and we refer to this transformation as a planar Galilean mapping.

Definition 4.2.1. *(Planar Galilean Mapping) A planar Galilean mapping is a linear transformation of* $\mathbf{u} = \begin{bmatrix} t & u & v & 1 \end{bmatrix}^T$*, representable as a nonsingular* 4×4 *matrix,*

$$\begin{bmatrix} t' \\ w'u' \\ w'v' \\ w' \end{bmatrix} = \begin{bmatrix} h_{11} & 0 & 0 & h_{14} \\ h_{21} & h_{22} & h_{23} & h_{24} \\ h_{31} & h_{32} & h_{33} & h_{34} \\ h_{41} & h_{42} & h_{43} & h_{44} \end{bmatrix} \begin{bmatrix} t \\ u \\ v \\ 1 \end{bmatrix}. \tag{4.10}$$

This matrix $\mathcal{H}$ is an inhomogeneous matrix that can be divided into an inhomogeneous part, i.e., the first row $\mathbf{h}^1$, and a homogeneous part, i.e., the second, third and fourth rows, $\mathbf{h}^2$, $\mathbf{h}^3$, and $\mathbf{h}^4$, respectively. Unlike the planar homography, this mapping does not form a group, i.e., the product of two planar Galilean matrices is not, in general, a planar Galilean matrix.

To estimate the parameters of this mapping, the homogeneous and inhomogeneous parts can be computed separately. The Direct Linear Transformation Algorithm (see [HZ00]) can be used to estimate the homogeneous part of this matrix since,

$$\begin{bmatrix} u_i' \\ v_i' \\ w_i' \end{bmatrix} \times \begin{bmatrix} \mathbf{u}_i^T \mathbf{h}^2 \\ \mathbf{u}_i^T \mathbf{h}^3 \\ \mathbf{u}_i^T \mathbf{h}^4 \end{bmatrix} = 0. \tag{4.11}$$

An over-determined homogeneous system of equations can be constructed as,

$$\begin{bmatrix} \mathbf{0}^T & -w'_i\mathbf{u}_i^T & v'_i\mathbf{u}_i^T \\ w'_i\mathbf{u}_i^T & \mathbf{0}^T & -u'_i\mathbf{u}_i^T \\ -v'_i\mathbf{u}_i^T & u'_i\mathbf{u}_i^T & \mathbf{0}^T \end{bmatrix} \begin{bmatrix} \mathbf{h}^2 \\ \mathbf{h}^3 \\ \mathbf{h}^4 \end{bmatrix} = 0. \qquad (4.12)$$

and the solution can be found using SVD (see Section 4.1 in [HZ00] for further details). For the inhomogeneous part, the following linear system of equations can be solved using least squares,

$$\left[\mathbf{u}_i^T\right]\mathbf{h}^1 = t'_i. \qquad (4.13)$$

4.3 Two View Geometry

In this section, we study the geometry of a pair of Galilean cameras. The cameras may move in different directions at different velocities. The coordinates of the corresponding projections in first and second camera are $\mathbf{u} = (t, uw, vw, w)^T$ and $\mathbf{u}' = (t', u'w', v'w', w')^T$ respectively. The imaged coordinates in the two cameras are $\mathbf{u} = \mathcal{M}\mathrm{U}$ and $\mathbf{u}' = \mathcal{M}'\mathrm{U}$. This pair of equations may be rewritten as

$$\mathcal{A}_G\mathbf{g} = 0, \text{ where } \mathcal{A}_G = \begin{bmatrix} m_{11} & 0 & 0 & 0 & m_{15} - t & 0 & 0 \\ m_{21} & m_{22} & m_{23} & m_{24} & m_{25} & u & 0 \\ m_{31} & m_{32} & m_{33} & m_{34} & m_{35} & v & 0 \\ m_{42} & m_{42} & m_{43} & m_{44} & m_{45} & 1 & 0 \\ m'_{11} & 0 & 0 & 0 & m'_{15} - t' & 0 & 0 \\ m'_{21} & m'_{22} & m'_{23} & m'_{24} & m'_{25} & 0 & u' \\ m'_{31} & m'_{32} & m'_{33} & m'_{34} & m'_{35} & 0 & v' \\ m'_{41} & m'_{42} & m'_{43} & m'_{44} & m'_{45} & 0 & 1 \end{bmatrix}, \tag{4.14}$$

and m_{ij} are the elements of $\mathcal{M}$ and $\mathbf{g} = [T, X, Y, Z, 1, -w, -w']^T$. Since $\mathcal{A}_G$ in the homogeneous system of Equation 4.14 is a 8×7 matrix, it must have a rank of at most six for a non-trivial solution to exist. As a result, any 7×7 minor must have a zero determinant. There are eight different ways to choose the 7×7 minor to solve the system, but only two interesting variations. The first selection uses both rows containing the temporal indices (t', t) and five rows containing the spatial indices (u, v, u', v'). The second selection uses one row containing the temporal indices and six rows containing the spatial indices. As in [GH97], $\det(\mathcal{A}_{Gi}) = 0$ will produce the fundamental polynomial that has interaction terms but no squared terms. Hence, there are exists a 6×6 matrix called the fundamental matrix of uniform motion such that

$$(t'u', t'v', t', u', v', 1)\mathbf{\Gamma}(tu, tv, t, u, v, 1)^T = 0. \tag{4.15}$$

However, in all of the eight variations (of different minors), nine interaction terms do not exist. There are a total of 36 possible interaction terms however of these only 27 appear.

Definition 4.3.1. *(Galilean Fundamental Matrix) If* $\mathbf{u}$ *and* $\mathbf{u}'$ *are video points corresponding to the same worldline under two Galilean cameras, there exists a* 6×6 *matrix* Γ *such that,*

$$\begin{pmatrix} t'u' \\ t'v' \\ t' \\ u' \\ v' \\ 1 \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} 0 & 0 & 0 & f_1 & f_2 & f_3 \\ 0 & 0 & 0 & f_4 & f_5 & f_6 \\ 0 & 0 & 0 & f_7 & f_8 & f_9 \\ f_{10} & f_{11} & f_{12} & f_{13} & f_{14} & f_{15} \\ f_{16} & f_{17} & f_{18} & f_{19} & f_{20} & f_{21} \\ f_{22} & f_{23} & f_{24} & f_{25} & f_{26} & f_{27} \end{pmatrix} \begin{pmatrix} tu \\ tv \\ t \\ u \\ v \\ 1 \end{pmatrix} = 0.$$

$\mathbf{\Gamma}$ can be written more compactly as,

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{0} & \Delta\mathbf{F}' \\ \Delta\mathbf{F} & \mathbf{F}_{00} \end{pmatrix}, \tag{4.16}$$

where $\mathbf{F}_{00}$ is the fundamental matrix between the image in the first video at time $t = 0$ and the image in the second video at time $t' = 0$, and $(\Delta\mathbf{F}, \Delta\mathbf{F}')$ are matrices that capture information about the velocity of each camera as will be seen presently.

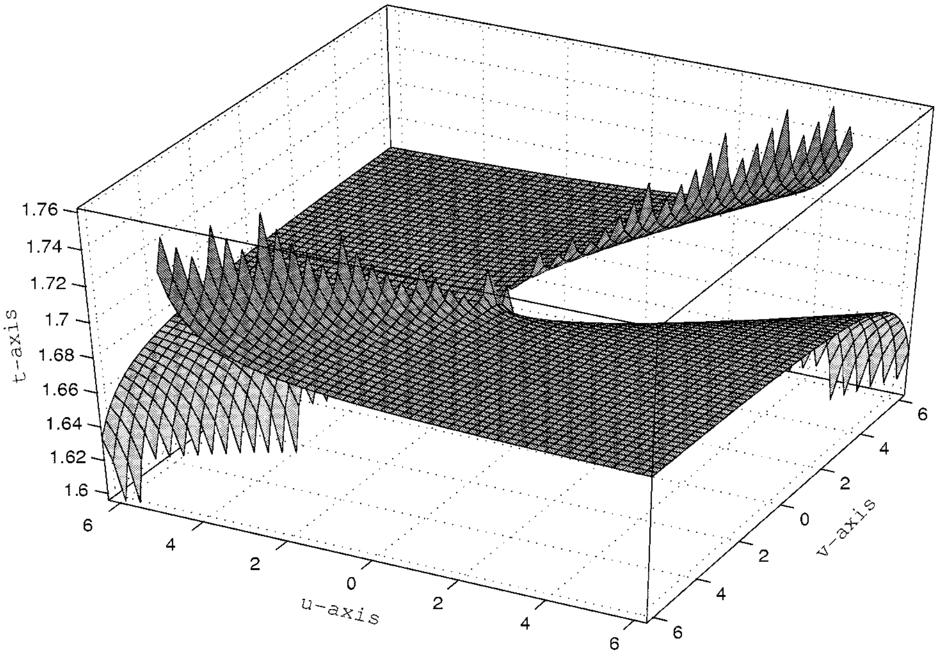


Figure 4.2: Epipolar surface. The spacetime point in the second video corresponding to a spacetime point in the first video must lie on this surface.

4.3.1 Epipolar Geometry

Unless there is zero motion, no epipoles (single image points of the opposite camera center) in the usual sense exist, and inn general, epipolar lines (or curves) in the usual sense do not exist either. Instead, there are epipolar surfaces in one camera corresponding to a point in the other camera. These surfaces are defined by setting a spatiotemporal point in one camera, i.e., given (t', u', v') and applying the Galilean fundamental matrix. The surface is defined by the

$$s_1tu + s_2tv + s_3t + s_4u + s_5v + s_6 = 0$$

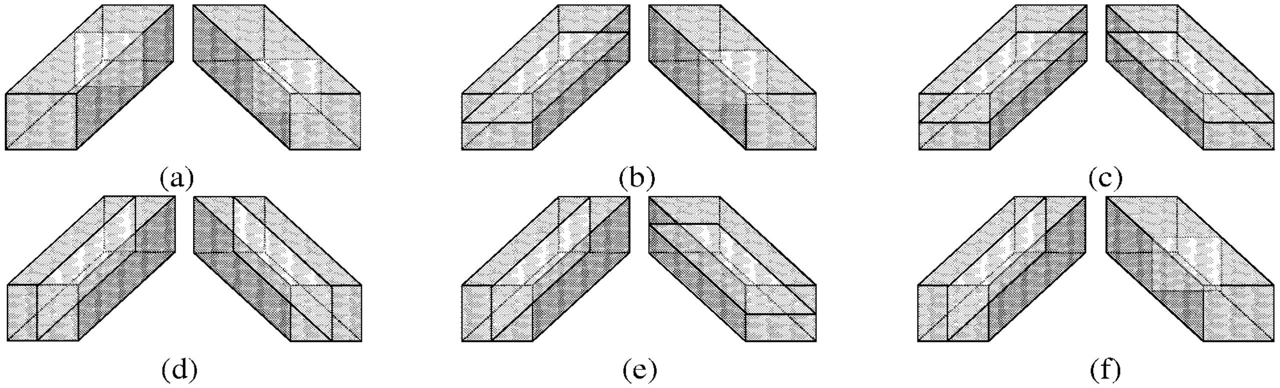


Figure 4.3: Specializations. "Fundamental" matrices can be recovered between (a) a pair of perspective images, (b) an EPI and a perspective images, (c) a pair of EPI images , (d) a pair of LP images, (e) a LP and EPI images, and (f) a LP and a perspective images.

where $\mathbf{s} = [s_1, \dots s_6]$ is computed as $\mathbf{s} = [t'u', t'v', t', u', v', 1]\mathbf{\Gamma}$. This surface is plotted in Figure 4.2, and is ruled, since the intersection with each time plane is a line (corresponding to the classic epipolar line of that image).

4.4 Specializations

Several specializations can be directly derived from the constraints described in this paper. The different specializations are shown in Figure 4.3(a), (b), and (c) for the original fundamental matrix, the orthoperspective fundamental matrix, and the linear pushbroom fundamental matrix, respectively. Similarly, it is straightforward to recover fundamental matrices (and planar transformations) for the configurations in Figure 4.3(d), (e), and (f).

4.4.1 Between Perspective Images

The classic fundamental matrix between two uncalibrated perspective images was derived independently by Faugeras in [Fau92] and Hartley in [Har92]. For corresponding points, this singular 3×3 matrix satisfies the constraint $[u', v', 1]\mathbf{F}[u, v, 1] = 0$. This matrix can be directly recovered from the Galilean fundamental matrix. For the (t', t) pair image we can recover the fundamental matrix $\mathbf{F}_{t't}$ by partially collapsing $\mathbf{\Gamma}$ and plugging in the values of (t', t). Thus,

$$\mathbf{F}_{t't} = \begin{pmatrix} f_1t' + f_{10}t + f_{13} & f_2t' + f_{11}t + f_{14} & f_3t' + f_{12}t + f_{15} \\ f_4t' + f_{16}t + f_{19} & f_5t' + f_{17}t + f_{20} & f_6t' + f_{18}t + f_{21} \\ f_7t' + f_{22}t + f_{25} & f_8t' + f_{23}t + f_{26} & f_9t' + f_{24}t + f_{27} \end{pmatrix},$$

or simply,

$$\mathbf{F}_{t't} = \Delta\mathbf{F}t + \Delta\mathbf{F}'t' + \mathbf{F}_{00}. \tag{4.17}$$

Thus, $\Delta\mathbf{F} + \mathbf{F}_{00}$ is the fundamental matrix between the image in the first video at time $t = 0$ and the second video at time $t' = 1$ and $\Delta\mathbf{F}' + \mathbf{F}_{00}$ is the fundamental matrix between the image in the first video at time $t = 1$ and the second video at time $t' = 0$. We can infer some interesting properties from Equation 4.17.

Theorem 4.4.1. *(Fundamental Boost Matrix) The matrices $(\Delta\mathbf{F}, \Delta\mathbf{F}')$ are rank-2 matrices. As a result, the rank of $\mathbf{\Gamma}$ is at most 5.*

Proof. A fundamental matrix $\mathbf{F}$ can be decomposed into $\mathbf{K}'^{-T}\mathbf{R}[-\mathbf{RC}]_\times\mathbf{K}^{-1}$. If the second camera is displaced by $\Delta\mathbf{C}$ then the fundamental matrix becomes,

$$\mathbf{K}'^{-T}\mathbf{R}[-\mathbf{R}(\mathbf{C}-\Delta\mathbf{C})]_\times\mathbf{K}^{-1} = \mathbf{K}'^{-T}\mathbf{R}[-\mathbf{RC}]_\times\mathbf{K}^{-1} - \mathbf{K}'^{-T}\mathbf{R}[\mathbf{R}\Delta\mathbf{C}]_\times\mathbf{K}^{-1} = \mathbf{F}_{00} + \Delta\mathbf{F} = \mathbf{F}_{10}$$

Since $[\Delta\mathbf{C}]_\times$ is a skew-symmetric matrix, it follows that $\Delta\mathbf{F}$ is a rank 2 matrix. The left 6×3 submatrix or the upper 3×6 submatrix of $\mathbf{\Gamma}$ are therefore both rank 2 matrices, and $\mathbf{\Gamma}$ has a rank of at most 5. □

4.4.2 Between Linear Pushbroom Images

If the camera motion satisfies the conditions described in [GH97], linear pushbroom images can be recovered from a vertical slice of the video volume. Between two such images, the LP fundamental matrix was derived by Gupta and Hartley in [GH97]. The relationship captured by this matrix is expressed as $(t'v', t', v', 1)\mathbf{F}_{u'u}(tv, t, v, 1)^T = 0$. Thus, given (u, u'), this 4×4 matrix too can be derived from Γ as follows:

$$\mathbf{F}_{u'u} = \begin{pmatrix} 0 & 0 & f_5 & f_4 u + f_6 \\ 0 & 0 & f_2 u' + f_8 & f_7 u + f_9 + (f_1 u + f_3) u' \\ f_{17} & f_{18} + f_{16} u & f_{20} & f_{19} u + f_{21} \\ f_{11} u' + f_{23} & f_{22} u + f_{24} + (f_{10} u + f_{12}) u' & f_{14} u' + f_{26} & f_{25} u + f_{27} + (f_{13} u + f_{15}) u' \end{pmatrix}. \tag{4.18}$$

It can be observed that the structure of the matrix is the same as the one derived in [GH97].

4.4.3 Between Epipolar Plane Images

Epipolar plane images were defined by Bolles *et al.* in [RM87] as the collection of epipolar lines that correspond to one epipolar plane in the world. If lateral camera motion is assumed, then we can recover the fundamental matrix between two EPIs. In this case it would have a similar form to the LP fundamental matrix,

$$\mathbf{F}_{v'v} = \begin{pmatrix} 0 & 0 & f_1 & f_2 v + f_3 \\ 0 & 0 & f_2 v' + f_7 & f_8 v + f_9 + (f_5 v + f_6) v' \\ f_{10} & f_{12} + f_{11} v & f_{13} & f_{14} v + f_{15} \\ f_{16} v' + f_{22} & f_{23} v + f_{24} + (f_{17} v + f_{18}) v' & f_{19} v' + f_{25} & f_{26} v + f_{27} + (f_{20} v + f_{21}) v' \end{pmatrix}. \qquad (4.19)$$

4.4.4 Between a Pushbroom and a Perspective Images

Recently in [SS06], Khan *et al* derived the 4×3 perspective-orthoperspective fundamental matrix between a pushbroom image and a perspective image. The relationship captured by this matrix is expressed as $(t'v', t', v', 1)\mathbf{F}_{u't}(u, v, 1)^T = 0$. This matrix can also be directly derived from $\mathbf{\Gamma}$. Thus, given (u', t), we can compute

$$\mathbf{F}_{u't} = \begin{pmatrix} f_4 & f_5 & f_6 \\ f_1u' + f_7 & f_2u' + f_8 & f_3u' + f_9 \\ f_{16}t + f_{19} & f_{17}t + f_{20} & f_{18}t + f_{21} \\ f_{10}tu' + f_{22}t + f_{13}u' + f_{25} & f_{11}tu' + f_{23}t + f_{14}u' + f_{26} & f_{12}tu' + f_{24}t + f_{15}u' + f_{27} \end{pmatrix}. \quad (4.20)$$

or simply $\begin{pmatrix} \phi\Delta\mathbf{F}' \\ \phi(t\Delta\mathbf{F} + \mathbf{F}_{00}) \end{pmatrix}$, where $\phi = \begin{pmatrix} 0 & 1 & 0 \\ u' & 0 & 1 \end{pmatrix}$. Similarly, it is straightforward to recover $\mathbf{F}_{u'v}$ the fundamental matrices between linear pushbroom and an epipolar plane images, and $\mathbf{F}_{v't}$ between an epipolar plane and a perspective images.

4.4.5 Between an Epipolar Plane and a Perspective Images

The relationship captured by this matrix is expressed as $(t'u', t', u', 1)\mathbf{F}_{v't}(u, v, 1)^T = 0$. This matrix can also be directly derived from $\mathbf{\Gamma}$. Thus, given (v', t), we can compute

$$\mathbf{F}_{v't} = \begin{pmatrix} f_1 & f_2 & f_3 \\ f_4 v' + f_7 & f_5 v' + f_8 & f_6 v' + f_9 \\ f_{10} t + f_{13} & f_{11} t + f_{14} & f_{12} t + f_{15} \\ t(f_{16} v' + f_{22}) + f_{19} v' + f_{25} & t(f_{17} t v' + f_{23}) + f_{20} v' + f_{26} & t(f_{18} t v' + f_{24}) + f_{21} v' + f_{27} \end{pmatrix}. \qquad (4.21)$$

4.4.6 Between an Epipolar Plane and a Linear Pushbroom Images

The relationship captured by this matrix is expressed as $(t'u', t', u', 1)\mathbf{F}_{v'u}(tv, t, v, 1)^T = 0$. This matrix can also be directly derived from $\mathbf{\Gamma}$. Thus, given (v', u), we can compute

$$\mathbf{F}_{v'u} = \begin{pmatrix} 0 & 0 & f_2 & f_1 u + f_3 \\ 0 & 0 & f_5 v' + f_8 & u(v' f_4 + f_7) + f_6 v' + f_9 \\ f_{11} & f_{12} + f_{10} u & f_{14} & f_{13} u + f_{15} \\ f_{17} v' + f_{23} & u(v' f_{16} + f_{22}) + v' f_{18} + f_{24} & f_{20} v' + f_{26} & u(f_{25} + v' f_{19}) + v' f_{21} + f_{27} \end{pmatrix}. \tag{4.22}$$

4.5 Normalized Linear Algorithm

A linear algorithm can be used to estimate the parameters of $\mathbf{\Gamma}$. Equation 4.16 can be rewritten as the homogeneous system $A\gamma = 0$, where $\gamma = [f_1, \cdots, f_{27}]^T$ is a 27-vector, constructed from the non-zero elements of $\mathbf{\Gamma}$ and $A\gamma$ is a matrix constructed from spatiotemporal coordinates of corresponding points. If noiseless points correspond exactly to each other the rank of $A\gamma$ is 26 and the null-vector of $A\gamma$ corresponds to an estimate of γ. In the presence of measurement noise, the 27th singular value will be non zero. In that case, the singular vector corresponding to the smallest singular value of $A\gamma$ can be used as an estimate of γ. The rank constraints on $\mathbf{\Gamma}$ can be enforced post facto by setting the singular values of each of the submatrices of $\mathbf{\Gamma}$ to zero and reconstructing the matrix. Of course, as with other linear algorithms of this sort, to obtain good

Objective

Given $n \geq 26$ matches from corresponding video points, estimate the Galilean fundamental matrix $\mathbf{\Gamma}$ such that $\mathbf{p}'^T\mathbf{\Gamma}\mathbf{p} = 0$. **Algorithm**

1. **Normalization**: Normalize the coordinates through an appropriate scaling and translation.
2. **Linear Solution**: Perform singular value decomposition on $A\gamma$ and determine $\mathbf{\Gamma}$ by selecting the singular vector corresponding to the smallest singular value of $A\gamma$ and reconstructing a 6×6 matrix.
3. **Rank Constraint**: By setting their smallest singular values to zero, enforce the rank 2 constraint on each of the submatrices of $\mathbf{\Gamma}$.
4. **Denormalization**: Denormalize $\mathbf{\Gamma}$ according to the original scaling and translation.

Figure 4.4: A linear algorithm for estimating $\mathbf{\Gamma}$.

estimates it is important to appropriately normalize the data (see Section 4.4 of [HZ00]). Lastly, to obtain a meaningful solution, $A\gamma$ has to have a rank of more than 26. It is emphasized that to ensure the rank of $A\gamma$ is greater than 17, correspondences from different events of the same point must be used. For instance, if videolines of a static world point in the scene of length n are associated in both cameras, there are n^2 rows that should be added to $A\gamma$. A minimum of 26 such correspondences are required, e.g., 7 points seen across two frames of each video would give 28 correspondences.

4.5.1 Linear transformation and Constant Velocity

In this section we show how the new fundamental matrix between uniformly moving cameras can be applied in action recognition framework. In the same manner as the dissimilarity measure was obtained from the fundamental matrix between perspective images, the new dissimilarity measure can be obtained from the fundamental matrix between Galilean cameras. We describe a new metric that can match actions despite linear transformation in time (scaling and shifts). We show that this metric can also match actions despite constant velocity motion of the camera. This model works effectively for many applications, particularly when the pattern is of a short duration. It was found that the use of a linear model is also appropriate for *coarse* matching and synchronization. Given a model action and a test action, we can deduce whether the actions observed in both sequences were equivalent up to a linear temporal transformation.

A linear transformation of time can be expressed as,

$$t' = a_1 t + a_2, \tag{4.23}$$

where a_1 is a scaling and a_2 is a shift in time. In addition to differing rates of action execution, it is important to note that two cameras might have a different frame rate, and the starting points of the video in two cameras might also be shifted relatively in time. Furthermore, to remain stable despite constant velocity motion of the camera, we use the fundamental constraint of linear motion, Equation 4.16, between cameras moving independently with constant velocity. The relationship between points from the two sequences can be expressed as,

$$
\mathcal{A}_{\mathbf{T}}\gamma = \left[\begin{matrix}
u'_1t'_1u_1 & u'_1t'_1v_1 & u'_1t'_1 & v'_1t'_1u_1 & v'_1t'_1v_1 & v'_1t'_1 & t'_1u_1 & t'_1v_1 & t'_1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
u'_nt'_nu_n & u'_nt'_nv_n & u'_nt'_n & v'_nt'_nu_n & v'_nt'_nv_n & v'_nt'_n & t'_nu_n & t'_nv_n & t'_n
\end{matrix}\right.
$$

$$
\begin{matrix}
u'_1u_1t_1 & u'_1v_1t_1 & u'_1t_1 & u'_1u_1 & u'_1v_1 & u'_1 & v'_1u_1t_1 & v'_1v_1t_1 & v'_1t_1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
u'_nu_nt_n & u'_nv_nt_n & u'_nt_n & u'_nu_n & u'_nv_n & u'_n & v'_nu_nt_n & v'_nv_nt_n & v'_nt_n
\end{matrix}
$$

$$
\left.\begin{matrix}
v'_1u_1 & v'_1v_1 & v'_1 & u_1t_1 & v_1t_1 & t_1 & u_1 & v_1 & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
v'_nu_n & v'_nv_n & v'_n & u_nt_n & v_nt_n & t_n & u_n & v_n & 1
\end{matrix}\right]\gamma = 0, \qquad (4.24)
$$

where γ is a 27-dimensional vector and $\mathcal{A}_{\mathbf{T}}$ is a matrix constructed from spacetime image coordinates of the corresponding points. If points exactly correspond to each other, then the rank of $\mathcal{A}_{\mathbf{T}}$ is 26, otherwise, the 27^{th} singular value will be non zero. Thus, instead of estimating κ from the observation matrix associated with the original fundamental matrix, we construct this new observation matrix and use the ratio of the 27^{th} to the 1^{st} singular values as our measure of dissimilarity.

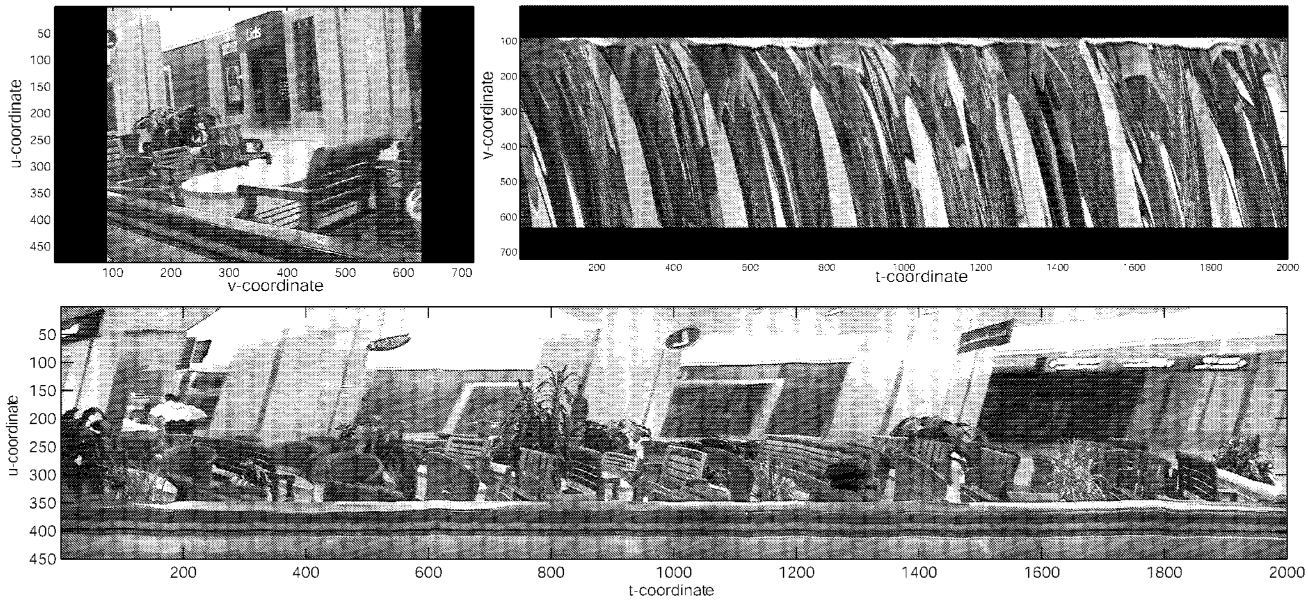
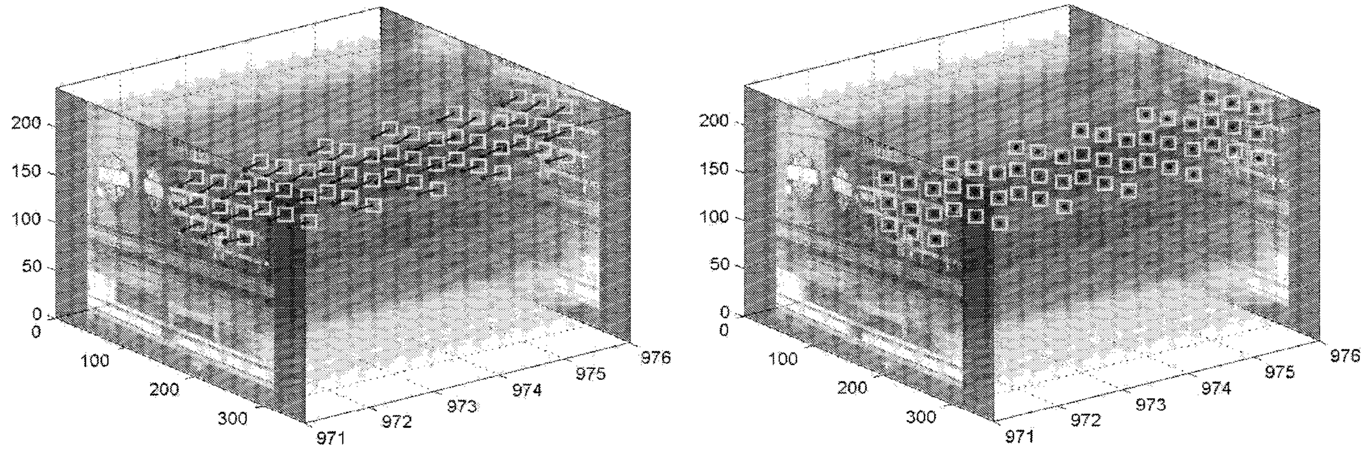


Figure 4.5: Three slices of a video taken from a Galilean camera. From left to right and top to bottom: perspective image, epipolar plan image and linear pushbroom image.



(a) (b)

Figure 4.6: Video points mapped using planar Galilean mapping. Yellow points indicate the position of the points in Sequence 1, black points indicate the position of (a) corresponding points in Sequence 2 and (b) corresponding points after warping.

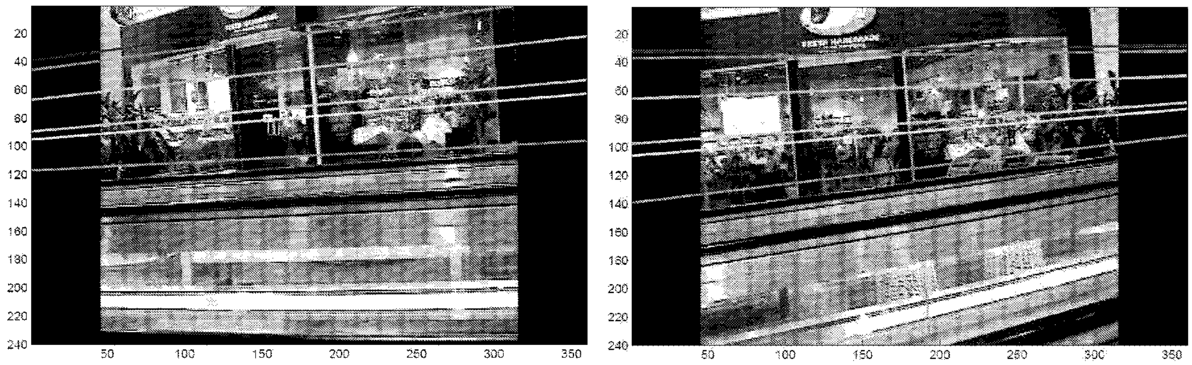


Figure 4.7: Recovering the fundamental matrix between frame 1 in sequence 1 and frame 20 in sequence 2 from the Galilean fundamental matrix. The Galilean fundamental matrix was computed from video points in six frames (971 to 976 in both sequences).

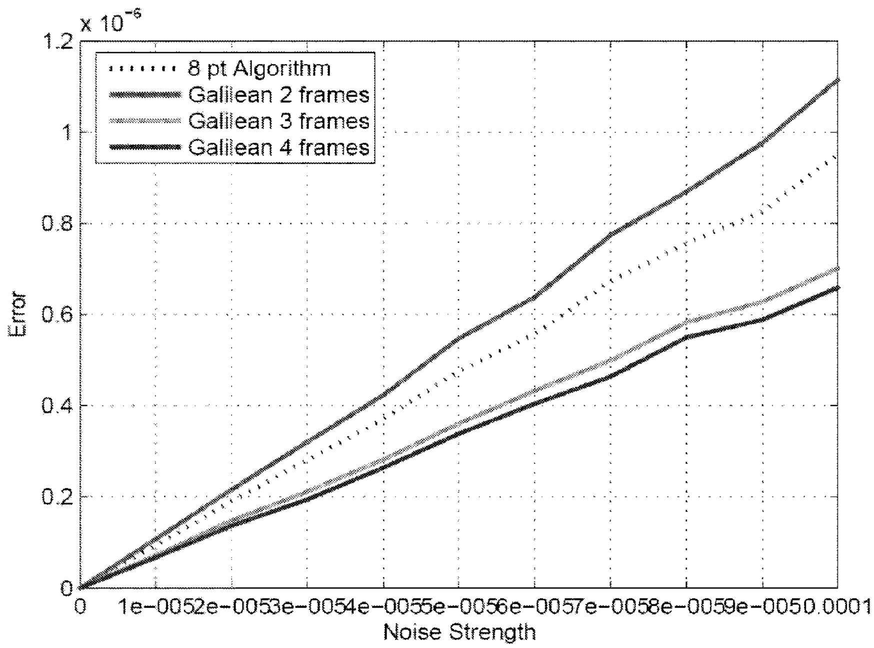


Figure 4.8: The accuracy of the fundamental matrix extracted from Galilean matrices computed from 2, 3 and 4 frames with respect to noise.

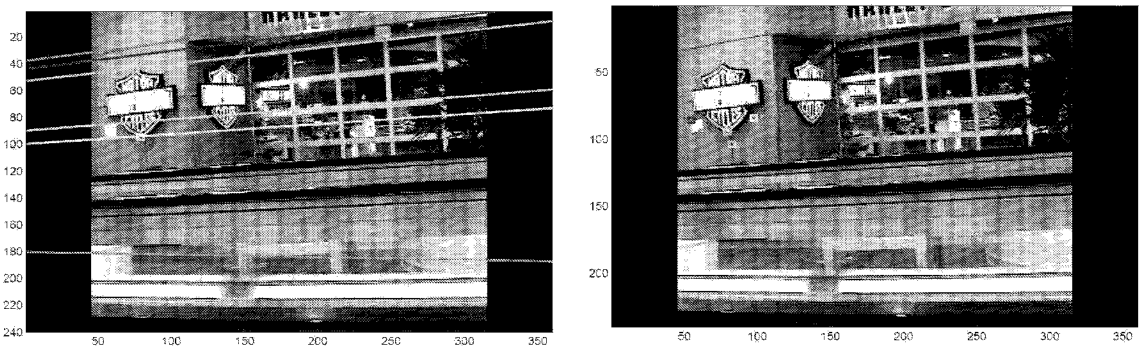


Figure 4.9: Frame 980 with the points inducing the epipolar curves in Figure 4.10 and Figure 4.11.

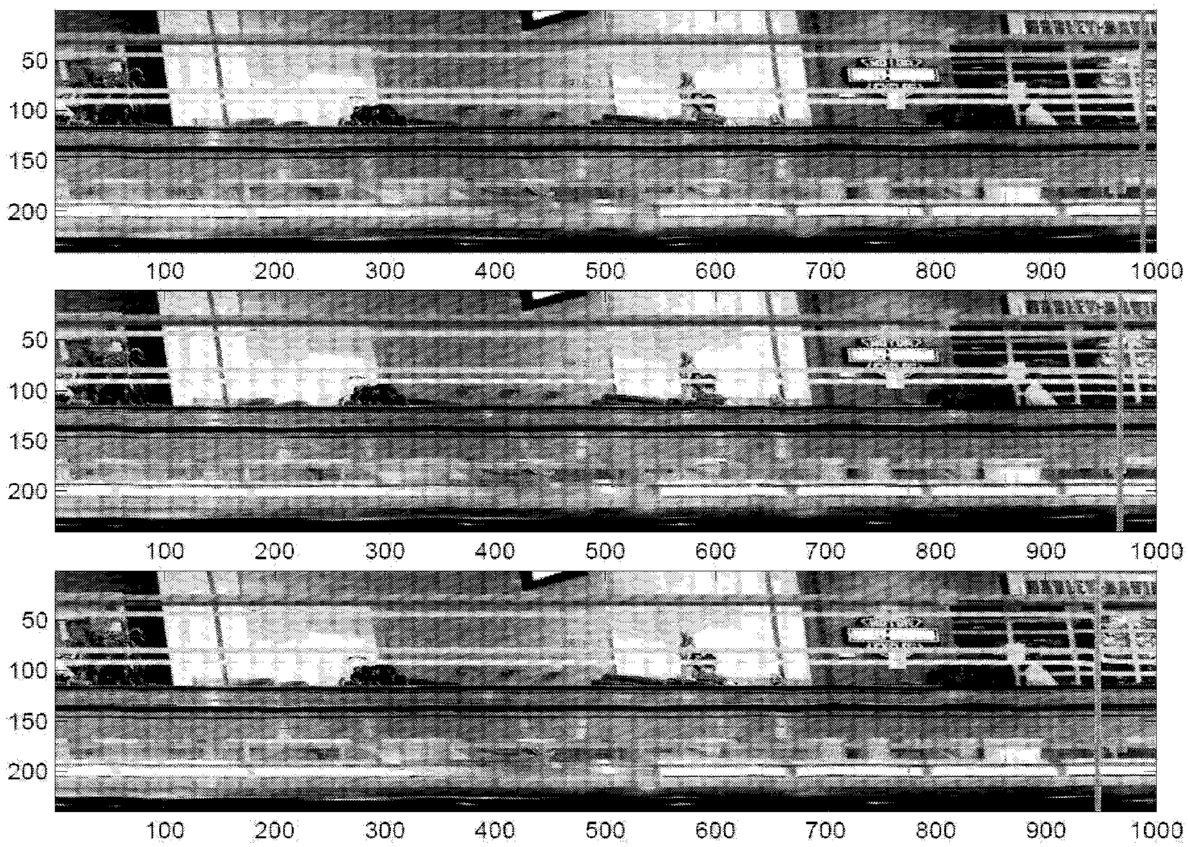


Figure 4.10: The epipolar curves induced by frames 960, 980 and 1000 of camera 2 on the pushbroom images of camera 1 corresponding to column 250.

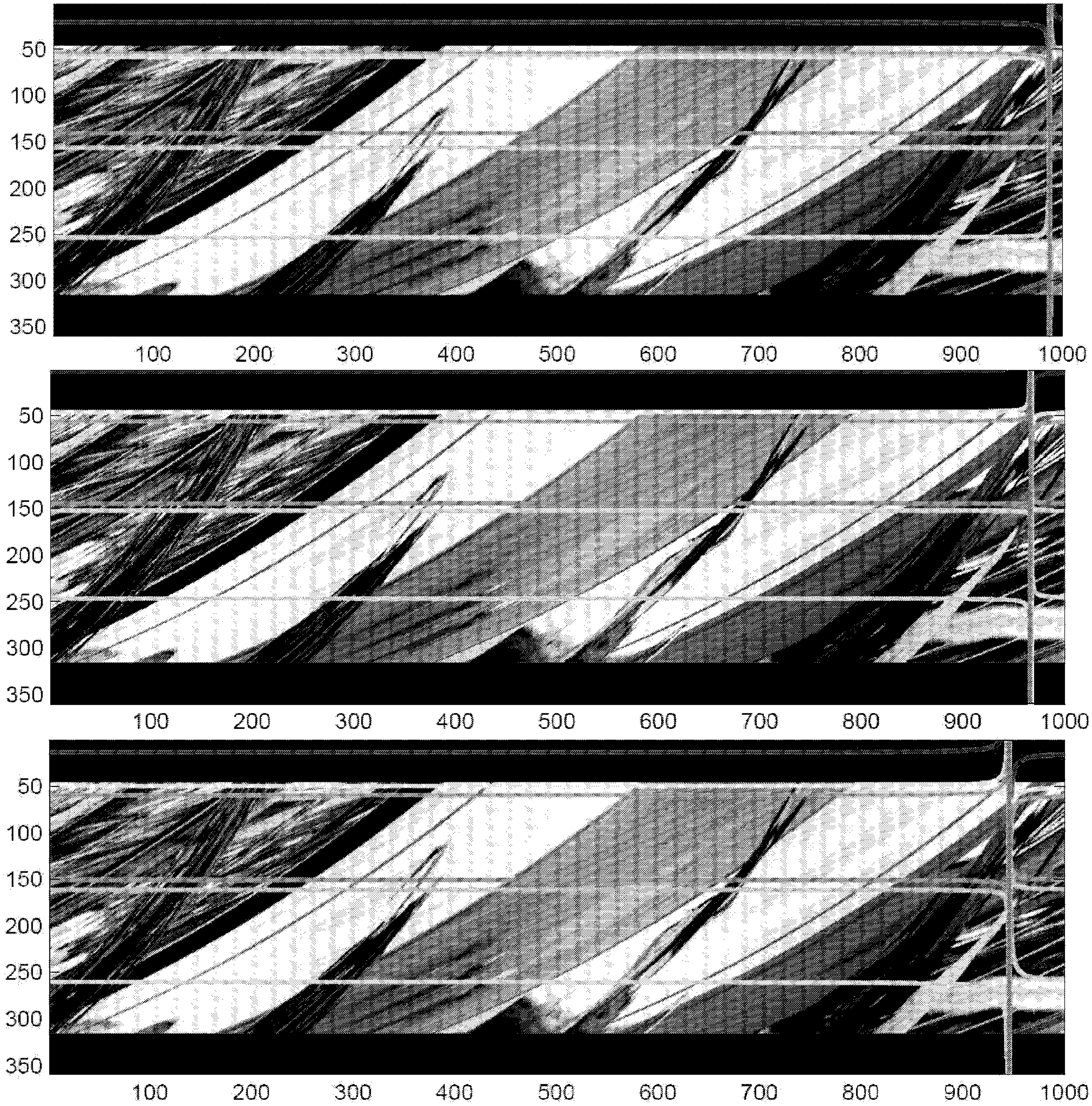
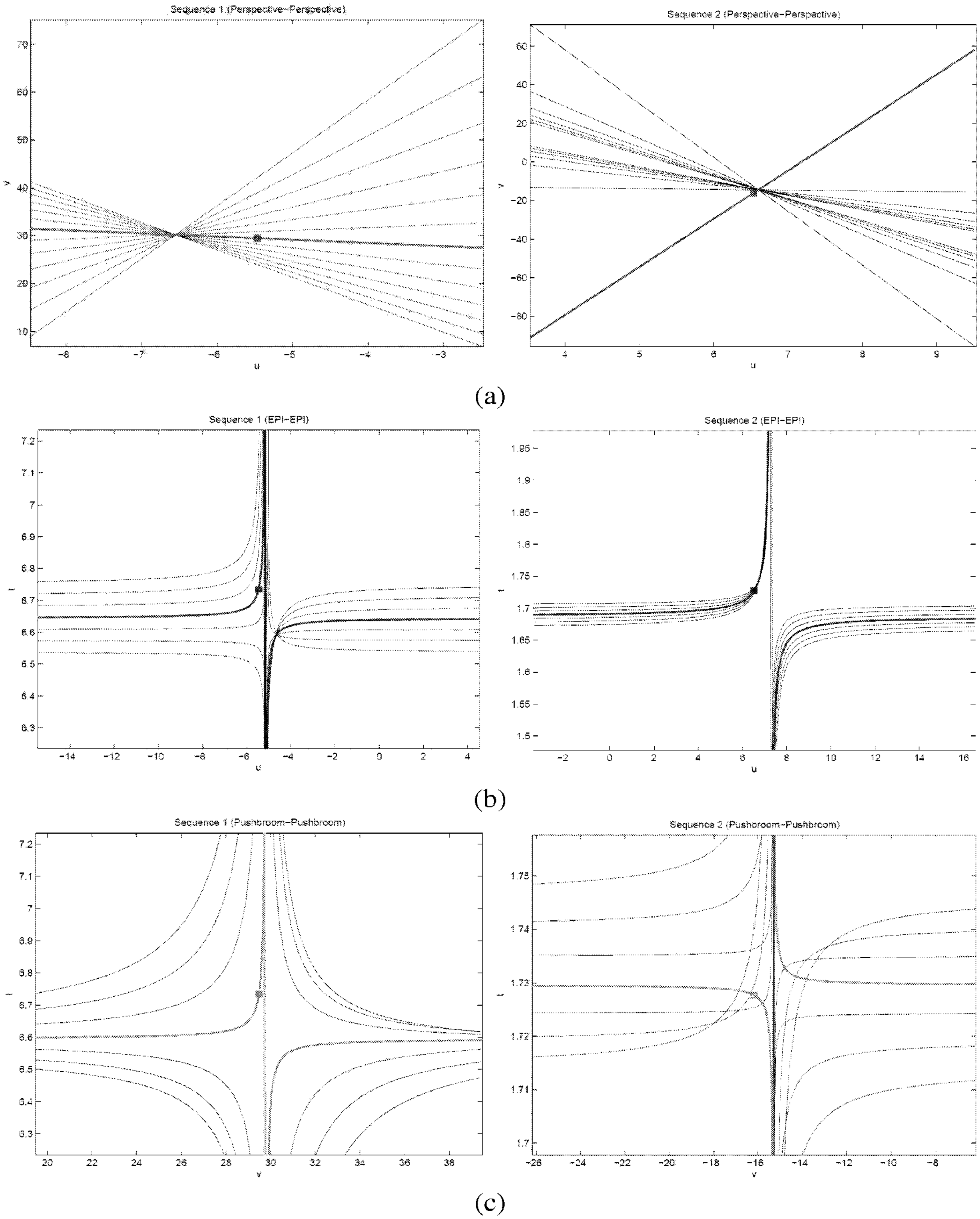


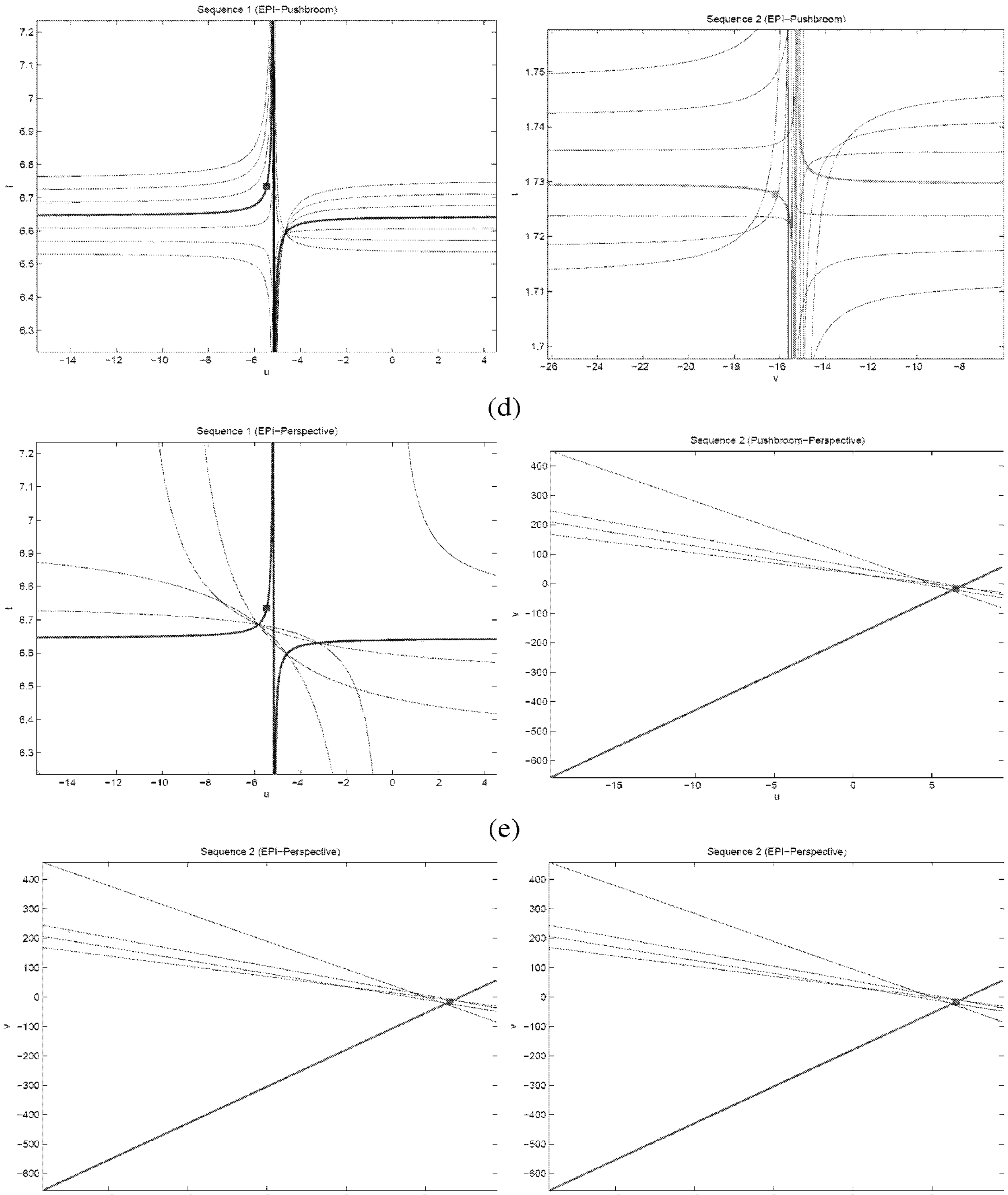
Figure 4.11: The epipolar curves induced by frames 960, 980 and 1000 of camera 2 on the EPI images of camera 1 corresponding to row 100.


Sequence 1 (Perspective-Perspective)
v
70
60
50
40
30
20
10
-8
-7
-6
-5
-4
-3
u
Sequence 2 (Perspective-Perspective)
v
60
40
20
0
-20
-40
-60
-80
4
5
6
7
8
9
u
Sequence 1 (EPI-EPI)
t
7.2
7.1
7
6.9
6.8
6.7
6.6
6.5
6.4
6.3
-14
-12
-10
-8
-6
-4
-2
0
2
4
u
Sequence 2 (EPI-EPI)
t
1.95
1.9
1.85
1.8
1.75
1.7
1.65
1.6
1.55
1.5
-2
0
2
4
6
8
10
12
14
16
u
Sequence 1 (Pushbroom-Pushbroom)
t
7.2
7.1
7
6.9
6.8
6.7
6.6
6.5
6.4
6.3
20
22
24
26
28
30
32
34
36
v
Sequence 2 (Pushbroom-Pushbroom)
t
1.75
1.74
1.73
1.72
1.71
1.7
-26
-24
-22
-20
-18
-16
-14
-12
-10
-8
v

(a)

(b)

(c)



(d)

(e)

(f)

Figure 4.12: Corresponding points and their epipolar curves. (a) Between two perspective images, (b) between two LP images, (c) between two EPIs, (d) between a LP image and EPI, (e) between LP image and a perspective, and (f) between an EPI and a perspective.

4.6 Results

4.6.1 Specializations

An experiment was conducted where two cameras were placed on a moving walkway at a distance of 8 feet, looking at different angles but both moving at approximately 2 miles per hour in the same direction. A pair of 1000 frame sequences were recorded at a resolution of 240×360 by two SONY HDV cameras (images were down sampled) and 22 video points were tracked across 6 frames in each of the two videos (frames 971 to 976 in both sequences). The motion of the cameras was not perpendicular to the optical axis of either camera. At 30 fps, the distance traversed by both cameras during this period was about 98 feet, and the distance traveled in between successive frames was approximately 1.173 inches. Three slices of this video are shown in Figure4.5. These points were used to estimate the Galilean fundamental matrix using the linear algorithm presented in this paper. To evaluate this estimate, different time slices of the video were analyzed using the different specializations of the Galilean fundamental matrix. A set of 6 points (different from the ones used during computation) were selected in both frames and the points and the epipolar lines of their correspondences between two perspective images are plotted in Figure 4.7. Despite the fact that the video's point correspondences were taken towards the end of the sequence and the frames in this figure were taken towards the beginning (frames 1 from video 1 and frame 20 from video 2), the fundamental matrix recovered is accurate. The epipolar curves induced by points in frames 960, 980, and 1000 onto the pushbroom image in video 1 generated from column 250 are shown in Figure 4.7. As the corresponding frames move, the asymptote translates in the LP image translates as well. Figure 4.9 shows frame 980 with the points used to plot the curves in Figures 4.7 and 4.11. The corresponding epipolar lines from the EPI are missing since it is difficult to find point correspondences on EPIs. Nine points lay on a plane and were used to compute the planar Galilean mapping shown in Figure 4.6. The figure shows the first sequence where the yellow boxes show the positions of the nine points in that sequence and the black points indicate the positions of the points in Sequence 2 (a) before and (b) after warping.

Simulations were also conducted to evaluate the accuracy of the fundamental matrix F obtained from Γ as the measurement noise strength was increased and the number of frames used to compute Γ was also increased. The fundamental matrix estimated using the eight point algorithm was also included for context. The evaluation was conducted by generating 12 random 3D points and two randomly placed Galilean cameras and computing an exact fundamental matrix from both projection matrices at time T=0 (generated using MATLAB code provided with [HZ00]). We then evaluated the estimated fundamental matrices with respect to increasing time duration and increasing measurement error (assumed to be normally distributed) using the distance measure provided by Zhang in Section 4.1 of [Zha98]. The experiment was conducted 500 times at each noise level and the average error was recorded. The result is shown in Figure 4.8. The epipolar curves induced by the different specializations of the estimated Γ are shown in Figure 4.12. The color coding is red, green and blue for perspective, EPI and LP images, respectively.

4.6.2 Application of the New Metric in Action Recognition

The following experiments quantitatively demonstrate that the new dissimilarity measure is stable to changes in viewpoint, anthropometry, and temporal behavior. Experiments were performed to evaluate each of these three properties in isolation, followed by experiments under all three simultaneously. Similarly to experiments described in Section 3.5.1, motion capture data was used to provide 3D data, which was projected onto 2D and used in all experiments. In the experiments, actions were observed from different locations in the upper hemisphere. Results are now analyzed and explained in detail.

4.6.2.1 Viewpoint

This set of experiments was performed on four actions, 'ballet', 'standing up', 'sitting down' and 'walking', testing the sensitivity of the metric with respect to noise and its behavior with respect to increase in number of frames. The experimental results are presented in the Figure 4.13. The pattern actions were all observed from a fixed viewpoint, and the azimuth and elevation were 30 and 25 degrees, respectively. The test actions were observed from a significantly distinct view angle, where the azimuth and elevation were 130 and 45 degrees, respectively. Six levels of noise, sampled from a zero-mean normal distribution with σ varying from 0.6 to 3.6, were added to the test actions. Twenty five samples were generated from at each noise strength and the mean at each noise level was recorded. The experiments showed κ to be robust to noise. As expected, estimates

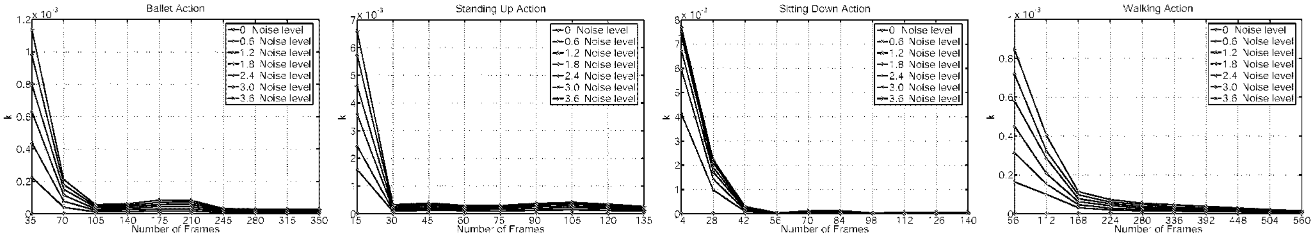


Figure 4.13: The measure is robust to changes in viewpoint (different markers differentiate the different noise levels). This figure shows how the proposed dissimilarity measure changes with respect to the level of noise and the view angle. Patterns of four actions were captured at the same view point, azimuth=30 and elevation=25 degrees, and test actions were observed by the stationary camera at view point corresponding to the azimuth=130 and elevation=45 degrees. Six levels of noise, sampled from zero mean normal distribution with σ varying from 0.6 to 3.6, were added to the 2D image coordinates. Regardless of the action, when the length of the action increases, the dissimilarity measure approaches zero. The $X-$ axis shows the number of frames, and the $Y-$axis shows the values of κ.

of κ become more reliable as the number of frames increases, and the number of frames after which κ is stable, varies from action to action, depending largely on the 'content' of the action.

4.6.2.2 Anthropometry

In this set of experiments we examined the performance of κ with respect to change in the anthropometry of the actor. The second row of Figure 3.5 shows the 'getting up' action under different anthropometric transformations. Figure 4.14 presents the experimental results. The pattern actions were observed from a view point with a fixed elevation of 60 degrees, while the azimuth was changed from 0 to 350 degrees. Similarly, the test action was observed from a view point with a fixed elevation of 30 degrees, while the azimuth was changed from 0 to 350 degrees. A 4×4 matrix $\mathcal{M}$ was randomly generated, and the whole action was transformed by $\mathcal{M}$. After 3D projective transformation, 3D points were projected onto image plane and distorted by six different

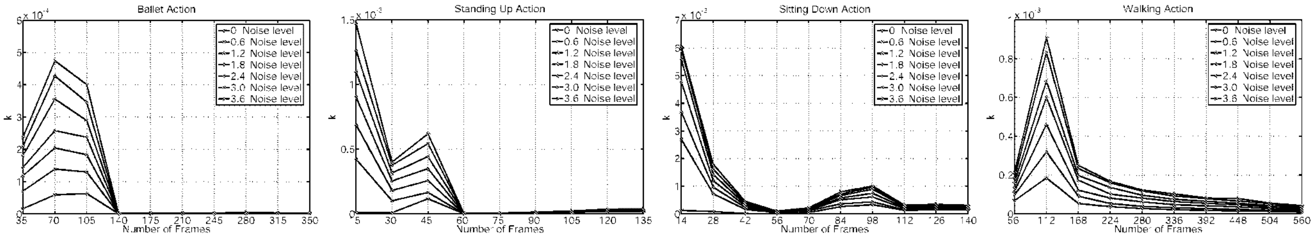


Figure 4.14: The proposed dissimilarity measure is stable to anthropometric distortion. The figure shows how κ changes with respect to the level of noise and the length of the action. Patterns of four actions were captured at the same view point, azimuth=30 and elevation=25 degrees, and test actions were observed by the moving camera at the different view point, azimuth=130 and elevation=25 degrees. Six different levels of noise, sampled from a zero-mean normal distribution with σ varying from 0.6 to 3.6, were added to the 2D image coordinates. Regardless of the action, when the length of the action increases, the dissimilarity measure approaches zero.

levels of noise. Noise parameters were the same as in the previous set of experiments. The results showed κ to be robust to noise, and estimates of κ become more reliable as the number of frames increases. As in the previous experiment, the number of frames after which κ is stable, varies from action to action and depends on the 'content' of the action.

4.6.2.3 Execution Rate

This set of experiments demonstrate the robustness to temporal distortion of actions with the results shown in Figure 4.15. Patterns of four actions were observed at a constant viewpoint, and the azimuth and elevation were 30 and 25 degrees, respectively. Test actions were observed at different view angles corresponding to the azimuth of 130 and elevation at 45 degrees. The test actions were distorted temporally by generating a pair (a_1, a_2) and by the same six levels of noise.

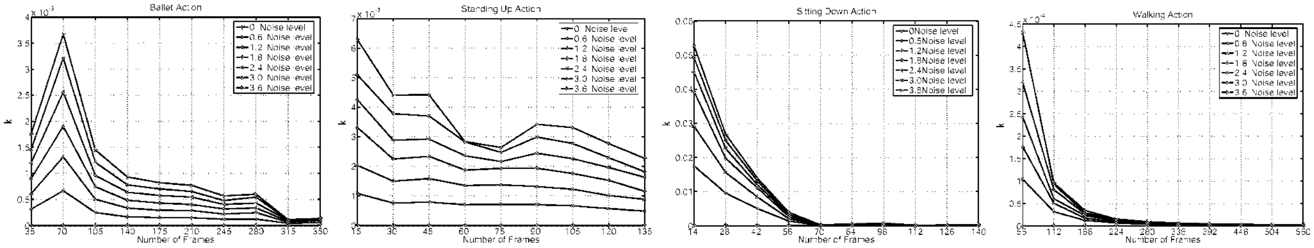


Figure 4.15: The proposed dissimilarity measure is stable to temporal distortion. The figure shows how the proposed dissimilarity measure changes with respect to the level of noise and the length of the action. Patterns of four actions were captured at the same view point, azimuth=30 and elevation=25 degrees, and test actions were observed by the moving camera at the different view point, azimuth=130 and elevation=25 degrees. Six different levels of noise, sampled from a zero-mean normal distribution with σ varying from 0.6 to 3.6, were added to the 2D image coordinates. Regardless of the action, when the length of the action increases, the dissimilarity measure is approaching zero.

As in the previous examples, increasing the duration of the actions increased the robustness to noise.

4.6.2.4 Simultaneous Distortion of Temporal Index, Viewpoint and Anthropometry

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

The first set of experiments was performed on synthetic data. The results presented in Figure 4.16 demonstrate the behavior of the dissimilarity measure, κ, with respect to all three types of transformations. Patterns of four actions, ‘ballet’, ‘standing up’, ‘sitting down’, and ‘walking’, were captured by the virtually moving camera with a fixed orientation, azimuth=30 and elevation=45 degrees. Test actions were captured by the virtually stationary camera at the view point

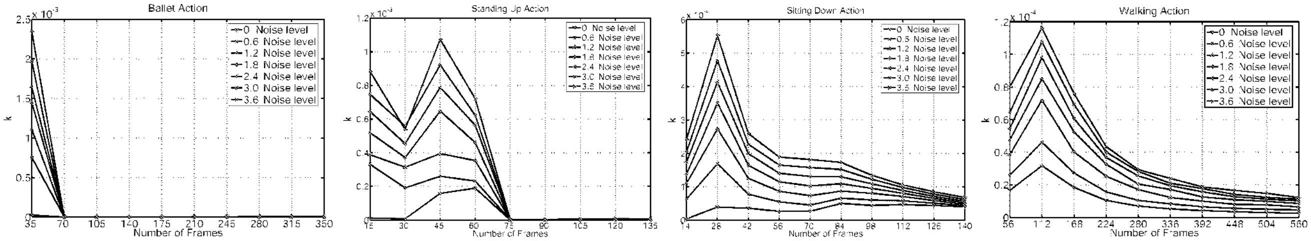


Figure 4.16: The dissimilarity measure is robust to view, anthropometric and temporal distortions. This figure shows how the dissimilarity changes with respect to the level of noise and the length of the action. From left to right, four figures correspond to 'ballet', 'standing up', 'sitting down', and 'walking actions'. Patterns of four actions were captured by the moving camera with a fixed orientation, azimuth=30 and elevation=45 degrees, and test actions were observed by the stationary camera at the view point with azimuth=130 and elevation=10 degrees. Six different levels of noise, sampled from the normal distribution with means from 0.6 to 3.6 and σ= 1, were added to the 2D image coordinates. When the length of the action increases, the dissimilarity approaches zero. The $X-$axis shows the number of frames, and the $Y-$axis shows the values of κ.

with azimuth=130 and elevation=10 degrees. Such orientations were chosen on purpose for reasonably good action observations. Similar to the previous experiments, six different levels of noise sampled from the normal distribution with means from 0.6 to 3.6 and σ=1 were added to the image coordinates. All results show robustness of κ with respect to noise. From Figure 4.16, it can be noted that approaching zero depends on length and type of the action.

The second set of experiments performed on both synthetic and real video. The synthetic video contained a walking actor and was about 570 frames long. The synthetic model action was a short fragment of 70 frames captured by a virtually moving camera with constant velocity. Real video was captured both indoor and outdoor. The model action was a cycle of walking action captured outdoor by a stationary camera. The test actions were walking action and cycling on a recumbent bike captured indoor by stationary camera. The walking action was performed by two actors walking on a treadmill, and cycling was performed by one actor. In our case, the camera

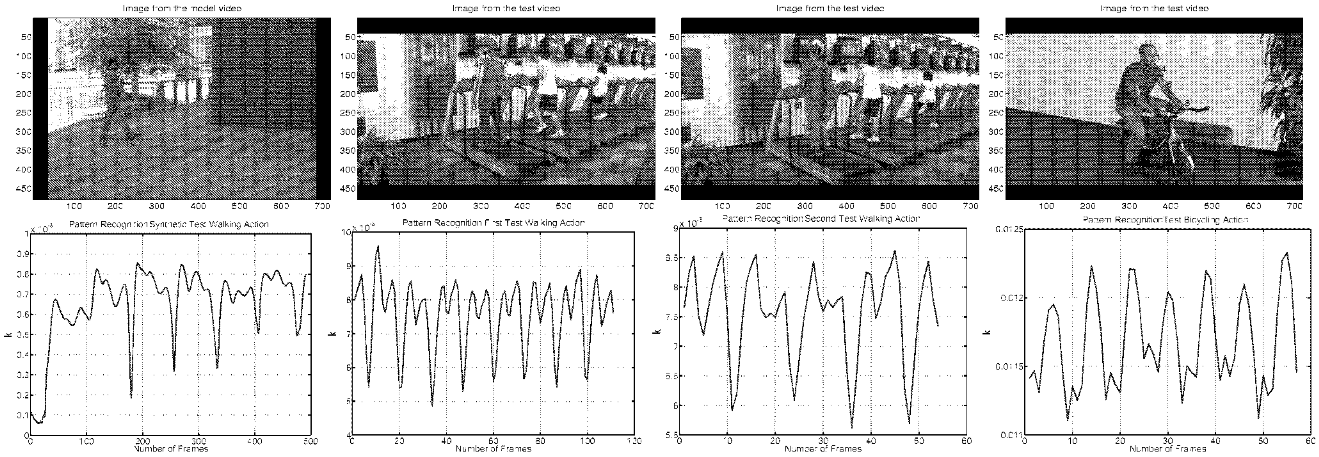


Figure 4.17: The first row shows images from real video. The left-most image corresponds to the model action, and the remaining three correspond to the test actions. The second row shows results of pattern detection. The left-most figure corresponds to recognition in synthetic video. The length of the model and test video was 70 and 564 frames, respectively. Two central figures shows detection of walking in real video containing walking actions. The model, and two test video were 42, 374, and 202 frames, respectively. In model and test video points on bodies were marked in each third frame. The right-most figure shows the result of detection of walking action in real video that did not contain any walking actions. The test video was 212 frames long. The values of local minima in the right-most figure are greater than ones in two central figures.

capturing query walking action can be interpreted as the camera moving in 3D along the direction of walking. The goal of the experiments was to determine whether the query actions contain a pattern action. The results are presented in Figure 4.17. The first row shows images from real video. The left-most image corresponds to the model action, and the remaining three correspond to the test actions. The model action was 42 frames long, and body joints were manually marked in each third frame. The test actions, two walking and bicycling actions, were the 374, 202, and 212 frames, respectively. As with the model video, points on the bodies were manually marked in each third frame of the test video. The second row shows the variation of κ as the pattern action was shifted in time over the duration of the test actions. The left-most figure shows the result of pattern detection in synthetic video, and other three figures show the results of pattern detection in real video. The right-most figure shows results of a video that did not contain any walking actions. Since that video contained an actor cycling, we do observe some periodicity. However, the values corresponding to each potential action occurrences (local minima) were greater than values corresponding to action occurrences in video contained walking action (see central figures).

While the above experiments determined only the location of action in test video (time translation), the final set of experiments determined the scale of temporal transformation. From all action occurrences in the previous experiment, only one occurrence was chosen in each video. Figure 4.18 demonstrates the results. The left-most figure from the top row shows the result obtained on synthetic video. The best match was detected when the scale of temporal transformation was one, and this coincides with the ground truth. The other two figures from the top row show the results obtained on real video. Compared to the model action, actors were walking slightly faster in both test video, and it was captured by the scale factor. In order to get the best match, actions from the test video were scaled to match the model action. Analyzing results from both synthetic and real video, it is easy to see that the global minima in the left-most figure is more distinct compared to the other two. This is attributed to noise, the length of the model action and our assumption, which is that we know the beginning point of action and do not know where action ends. As soon as the test fragment contains the action, κ becomes less sensitive to an increase in scale. This effect is still observable in synthetic video but to a lesser degree. The remaining three rows show the corresponding frames after synchronization between model action and test fragments.

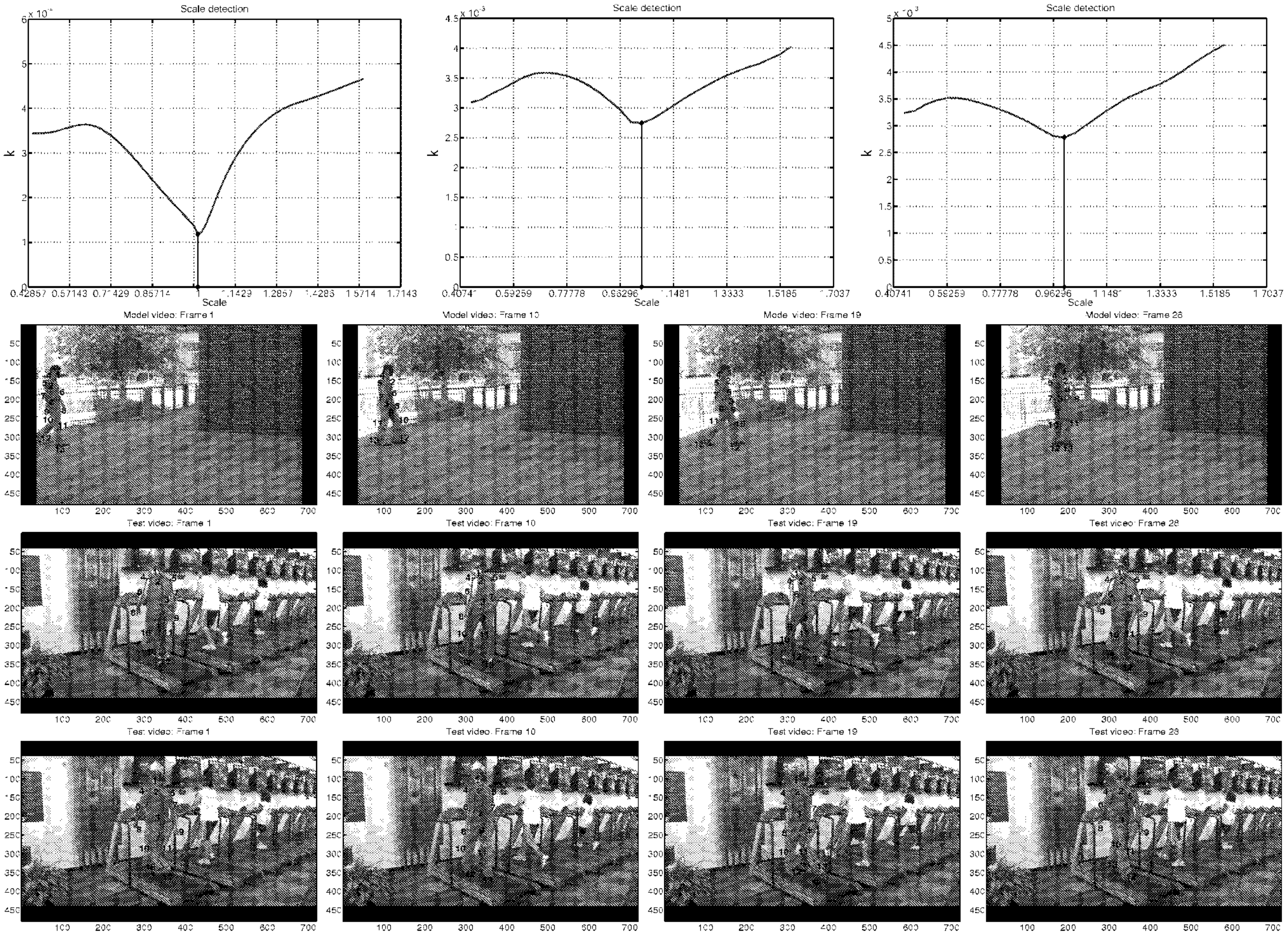


Figure 4.18: The top row shows the results of temporal scale detection. The left-most figure shows the result of detection in synthetic video. The best match corresponds to the point where the scale is one. The remain two figures show results of scale detection in real video. Since both actors were walking faster than in the model, the best matchings correspond to the scales, which are slightly greater than one. The remaining rows show the corresponding frames after synchronization. The second-top row shows frames from the model video, and others show frames from the test video.

4.7 Summary

A spacetime projection model for cameras moving at constant velocities in 3D space was derived. We investigated the relative geometry relating a pair of Galilean cameras in planar and general scenes and defined the planar Galilean mapping and the fundamental matrix that capture these relationships. Finally, we showed how three known fundamental matrices are specializations of this matrix, and could be readily recovered from the proposed fundamental matrix, and as a result, provide a unifying link between the classic fundamental matrix and the LP fundamental matrix. In addition, we described three new fundamental matrices that can also be recovered.

We have addressed the analysis of human action recognition in videos based on the presence of three key sources of distortion due to the viewpoint of observation, anthropometric proportion of actors, and differing rates of execution. The proposed approach matches actions based only on the imaged locations of anatomical landmarks across time. We demonstrated theoretically and then empirically that the algorithm based on the proposed dissimilarity measure is stable with respect to changes in all three distortions. During the experimentation, we examine each source of distortion in isolation, followed by an evaluation in the presence of simultaneous distortion and report the quantitative performance. In addition, we provide several qualitative examples demonstrating the applicability of the proposed approach.

CHAPTER 5
HUMAN BODY JOINT TRACKING USING GEOMETRICAL CONSTRAINTS

In previous chapters we have assumed that the landmark positions on a human body is given in each frame. In this chapter, it is assumed that only an initial correspondence between the first frames of a model and test video are known. Even with this assumption, human joint tracking in the remaining frames of a test video is a difficult task due to the continuous change of body parts appearance and large variations in action execution. We present a novel 2D model-based approach for tracking human body parts during articulated motion. The innovative use of an affine and epipolar line constraints allows us to reconstruct 2D motion of human body parts. Recovering motion trajectories significantly reduces a search space and provides us with good accuracy even in case of a significant change in a camera's viewpoint. The proposed method can be easily adopted for different articulated actions, does not require extensive tracking, and can be used with a single model. The performance and effectiveness of the developed tracking system is demonstrated in several different scenarios.

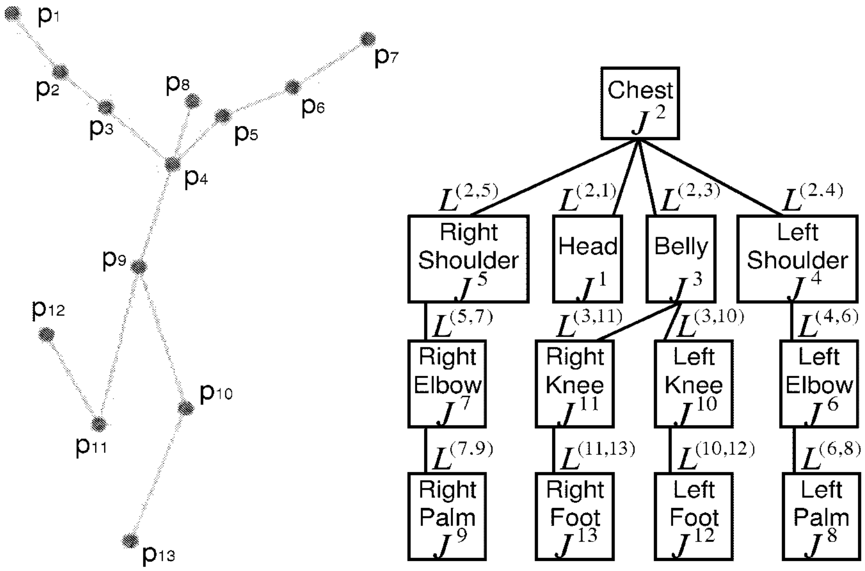
5.1 Human Model

If we consider a human joint as a point, there is not enough information to accurately localize this point in the image. In our framework, a model of a joint is not just a point on the human body, but a region around that point. The region around a body joint can provide us with color and edge information. From this point forward, any reference to a joint refers to a centroid of the region around the body joint. The detection and tracking of joints can be improved by imposing constraints on their mutual geometric coherence. In other words, the optimal location of the joints must preserve an appearance of the human body parts connecting them. Image regions corresponding to human body parts, or links, contain even more essential information than regions around body joints. Regions around body joints and regions corresponding to their links connecting joints can be perfectly embedded in a pictorial structure.

To facilitate the further explanation of a human model, we follow the definitions presented in Section 3.1. The pose and posture of an actor in terms of a set of points in 3D space is represented in terms of a set of 4-vectors $Q = \{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^n\}$, where $\mathbf{X}^k = (X^k, Y^k, Z^k, \Lambda)^\top$ are homogenous coordinates of the k^{th} joint. Each point represents the spatial coordinate of an anatomical landmark on the human body as shown in Figure 5.1. Points are connected by links corresponding to human body parts. Thus, a human body is represented as a pictorial structure defined as follow

$$P = (V, S),$$

where $V = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$ corresponds to landmarks on a human body in the image plane, with $\mathbf{x}^k = (a^k, b^k, \lambda)^\top$ and $S = \{L^{(k;j)} \mid k \neq j; k, j \in V\}$ corresponding to the links connecting



(a) (b)

Figure 5.1: (a) Point-based representation. (b) Pictorial Structure showing different body joints and their corresponding links.

landmarks. $\mathbf{X}^k$ and $\mathbf{x}^k$ are related by a 4×3 projection matrix $\mathcal{C}$, i.e., $\mathbf{x}^k = \mathcal{C}\mathbf{X}^k$. A known location of the k^{th} joint in the i^{th} frame of the model video, f_i^M, is denoted by $\mathbf{y}_i^k$. Similarly, a true unknown image location of the k^{th} joint in the j^{th} frame of the test video, f_j^T, is denoted by $\mathbf{x}_j^k$, and a candidate location of the k^{th} joint in f_j^T is denoted by $\hat{\mathbf{x}}_j^k$.

5.2 Joint Prediction

In this section we explain the details of the geometric constraints and how they are used to generate the predicted joint locations for the test video. A human body is modelled by a stick figure [Joh73] connecting 13 main landmarks points (head, neck, two shoulders, two elbows, two hands, belly point, two knees, and two feet). Figure 5.2 shows stick figures from two frames of a model video and corresponding frames of the test video. We assume that all the joint locations are available for the complete model video and the first frame of the test video. At this point we assume that the execution rate of the action in test video would be the same as that of the model. Hence, for every frame of the model video, we will predict the estimated stick figure for the corresponding frame in the test video. This mapping will capture the variations in viewpoints and the anthropometry of individuals. The variations in the rate and the style of execution will be addressed in the tracking phase. An affine constraint is first used to recover an initial estimate of the joint locations for the test sequence, followed by the refinement through the epipolar constraint. Both of these constraints are explained in the following two sections.

5.2.1 Affine Constraint

The geometric similarity between actions, depicted in model and test videos, is used to estimate the positions of landmarks in each frame of a test video.

In a real world an action can be represented as a sequence of 3D stick figures. Using this sequence of stick figures, a 3D trajectory can be recovered for any one of the landmarks. The points on these trajectories are used for the analysis of this section. Consider landmarks $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$ lying on the same plane in 3D and $\mathbf{X}_0$ that is off that plane. Connecting $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_4$ create a pyramid in 3D, and its volume is denoted by $V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}$. Consider volumes $V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}$, $V_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}$, observing from view 1, and $V'_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}$, $V'_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}$, observing from view 2. Considering the structure of a human action, the amount of out of plane motion is minimal as compared

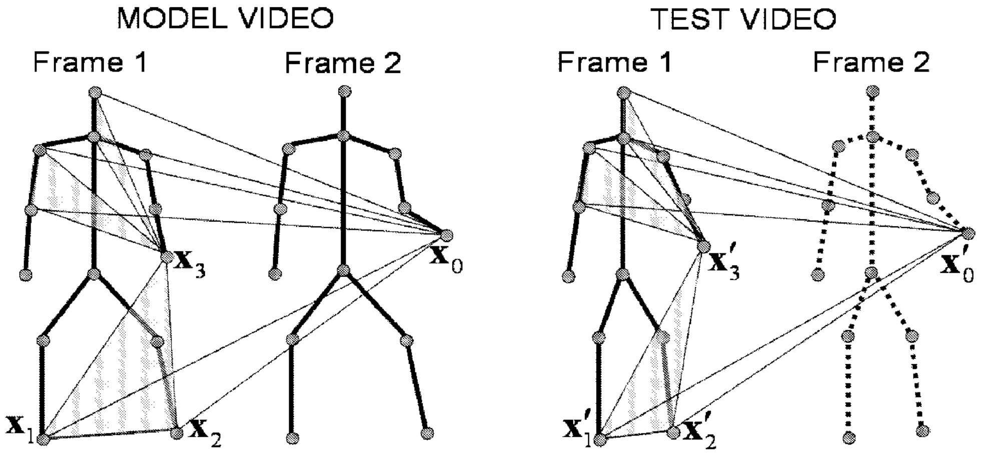


Figure 5.2: The locations, $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}'_1, \mathrm{x}'_2$ and x'_3, are known in the first frames of a model and test video. In each frame of a model video the location x_0 is also known, and x'_0 is required to recover in each frame of a test video. The unknown x'_0 can be computed from a system of quadratic equations, derived from ratios of areas of corresponding triangles in a model and test video, e.g. $\triangle(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3), \triangle(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_0), \triangle(\mathrm{x}'_1, \mathrm{x}'_2, \mathrm{x}'_3)$ and $\triangle(\mathrm{x}'_1, \mathrm{x}'_2, \mathrm{x}'_0)$.

to the distance from the view point. This allows us to impose affine transformation between two view points. It is known that under affine transformation the ratio of areas and ratio of volumes are preserved [HZ00], i.e.

$$\frac{V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}}{V_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}} = \frac{V'_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}}{V'_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}}, \tag{5.1}$$

These volumes can be computed as

$$\begin{aligned} V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4} &= \frac{S_{\mathbf{X}_4,\mathbf{X}_1,\mathbf{X}_2} h}{3}, V_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4} = \frac{S_{\mathbf{X}_4,\mathbf{X}_2,\mathbf{X}_3} h}{3}, \\ V'_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4} &= \frac{S'_{\mathbf{X}_4,\mathbf{X}_1,\mathbf{X}_2} h'}{3}, V'_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4} = \frac{S'_{\mathbf{X}_4,\mathbf{X}_2,\mathbf{X}_3} h'}{3}, \end{aligned}$$

where $S_{\mathbf{X}_4,\mathbf{X}_1,\mathbf{X}_2}$, $S_{\mathbf{X}_4,\mathbf{X}_2,\mathbf{X}_3}$, $S_{\mathbf{X}_4,\mathbf{X}_1,\mathbf{X}_2}$ and $S'_{\mathbf{X}_4,\mathbf{X}_2,\mathbf{X}_3}$ are the areas of base triangles, and h, h' are distances from $\mathbf{X}_0$ and $\mathbf{X}'_0$ to the planes of base triangles. Since, $\triangle(\mathbf{X}_4, \mathbf{X}_1, \mathbf{X}_2)$ and $\triangle(\mathbf{X}_4, \mathbf{X}_2, \mathbf{X}_3)$

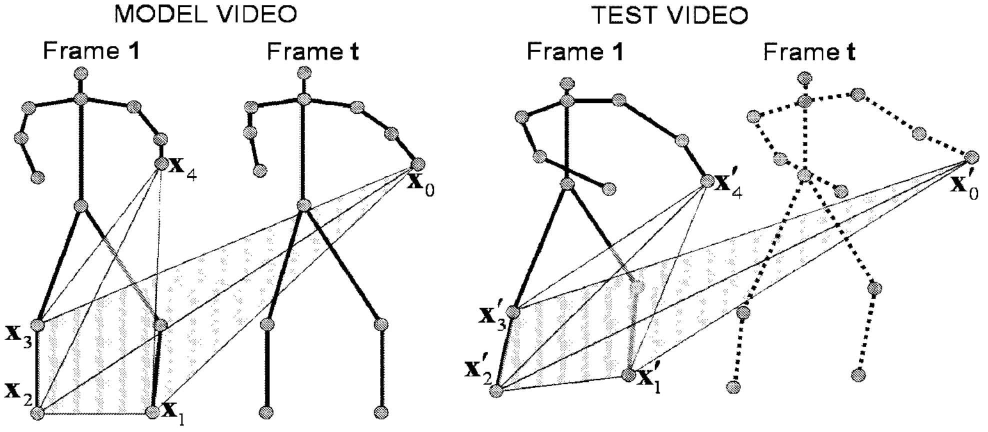


Figure 5.3: Fragment of the action in the model and test video. The locations of $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4, \mathrm{x}'_1, \mathrm{x}'_2, \mathrm{x}'_3$ and x'_4 and their correspondences are known; x_0 and x'_0 in frame t correspond to x_4 and x'_4 respectively. The location of x_0 is known, and x'_0 can be recovered from the areas of triangles. Triangles, shown in blue, are constructed from $\mathrm{x}_4, \mathrm{x}'_4$ and remain landmarks in the first frames. Triangles, shown in brown, are constructed in the same manner but $\mathrm{x}_0, \mathrm{x}'_0$ are used instead of $\mathrm{x}_4, \mathrm{x}'_4$.

lie on the same plane in 3D, the ratio of volumes in Eq.5.1 can be rewritten in terms of areas, projected onto image planes of both cameras

$$\frac{V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}}{V_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}} = \frac{S_{\mathbf{X}_4,\mathbf{X}_1,\mathbf{X}_2}}{S_{\mathbf{X}_4,\mathbf{X}_2,\mathbf{X}_3}} = \frac{S'_{\mathbf{X}_4,\mathbf{X}_1,\mathbf{X}_2}}{S'_{\mathbf{X}_4,\mathbf{X}_2,\mathbf{X}_3}} = \\ \frac{S_{\mathbf{x}_4,\mathbf{x}_1,\mathbf{x}_2}}{S_{\mathbf{x}_4,\mathbf{x}_2,\mathbf{x}_3}} = \frac{S'_{\mathbf{x}_4,\mathbf{x}_1,\mathbf{x}_2}}{S'_{\mathbf{x}_4,\mathbf{x}_2,\mathbf{x}_3}} = \kappa. \tag{5.2}$$

Figure 5.3 shows these *imaged* areas, where $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ and $\mathbf{x}_4$ are projections of $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ and $\mathbf{X}_4$ respectively. The ratio of volumes in Eq.5.1 and 5.2 can be also expressed in terms of different base areas and heights

$$V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4} = \frac{S_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_4} h_1}{3}, V_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4} = \frac{S_{\mathbf{X}_0,\mathbf{X}_3,\mathbf{X}_4} h_2}{3},$$

$$V'_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_2} = \frac{S'_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_4} h'_1}{3}, V'_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4} = \frac{S'_{\mathbf{X}_0,\mathbf{X}_3,\mathbf{X}_4} h'_2}{3}.$$

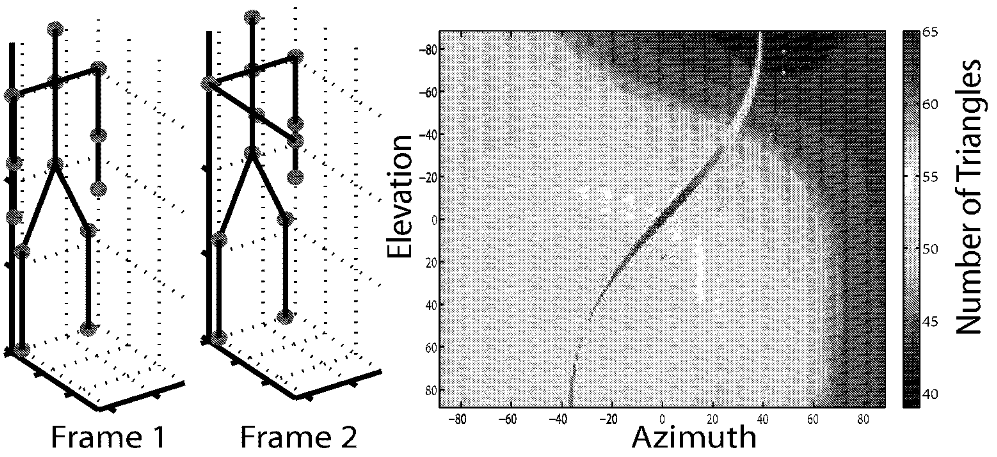
where h_1, h_2, h'_1 and h'_2 are distances from $\mathbf{X}_2$ to the planes of base triangles. Hence, the Eq.5.1 can be rewritten as

$$\frac{V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}}{V_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}} = \frac{S_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_4} h_1}{S_{\mathbf{X}_0,\mathbf{X}_3,\mathbf{X}_4} h_2} = \frac{S'_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_4} h'_1}{S'_{\mathbf{X}_0,\mathbf{X}_3,\mathbf{X}_4} h'_2} = \kappa. \tag{5.3}$$

We observed that for human articulated actions the ratio of volumes in Eq.5.3 can be approximated by areas, projected onto image planes

$$\frac{V_{\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_4}}{V_{\mathbf{X}_0,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4}} = \frac{S_{\mathbf{x}_4,\mathbf{x}_1,\mathbf{x}_2}}{S_{\mathbf{x}_4,\mathbf{x}_2,\mathbf{x}_3}} = \kappa \approx \frac{S_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_4}}{S_{\mathbf{x}_0,\mathbf{x}_3,\mathbf{x}_4}} \approx \frac{S'_{\mathbf{x}_0,\mathbf{x}_3,\mathbf{x}_4}}{S'_{\mathbf{x}_0,\mathbf{x}_3,\mathbf{x}_4}}. \tag{5.4}$$

The results of the following experiment validate the approximation in Eq.5.4. Using statistical anthropometric measurements from [Bai82], a virtual actor was created in 3D, Figure 5.4(a) shows two different postures of that actor in two consecutive frames. The distinct positions of the right arm provides with the significant change in depth that may occur in articulated actions. The model was observed from fixed view with zero elevation and azimuth. The test was observed from angle with elevation and azimuth varying independently from -88 to 88 degrees. According to the proposed approach, 65 triangles were constructed, using points in each video. Triangles with zero area were discarded. In the ideal affine case, the ratio of ratios must be one, but since, there is a large variation in the test camera view, the ratio of areas is not preserved. The goal of the experiment



(a) (b)

Figure 5.4: (a) Two different postures. The position of the right hand provides with the maximum depth change corresponding to the arm. (b) The relation between test camera view and ratio of ratios that fall in [0.95, 1.05].

was to determine the number of ratio of ratios that fall in the range from 0.95 to 1.05. The Figure 5.4(b) shows that there are more than 40 out of 64 ratio of ratios in the most of the cases that fall into that range.

Since under an affine transformation the ratio of areas remains the same [HZ00], this invariance is used to predict the unknown location of body joints in all the frames of the test video. For instance, let us predict the location of the left hand in the second frame of the test video using the proposed invariance of the ratio of triangular areas. Figure 5.2 shows the models of the two videos with known points and the unknown point $\mathbf{x}'_0$ corresponding to the left hand in second frame. All of these points are 2-dimensional vectors with x and y-coordinates defined in the image space. We apply the first constraint to determine $\mathbf{x}'_0$ using the areas from a pair of triangles $\triangle(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ and $\triangle(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_0)$ from the model video and another pair of $\triangle(\mathbf{x}'_1, \mathbf{x}'_2, \mathbf{x}'_3)$ and $\triangle(\mathbf{x}'_1, \mathbf{x}'_2, \mathbf{x}'_0)$ from the test video. Note that Figure 5.2 shows the two frames side by side just for illustration purposes.

In reality, these triangles can be imagined as the projection of one on top of another in the image plane. Let $S_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2}$ be the area of the corresponding triangle, hence, for the chosen pair of triangles, the invariance of the ratio of areas between the model and test video is presented as

$$\frac{S_{\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2}}{S_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3}} = \kappa \approx \frac{S_{\mathbf{x}'_0,\mathbf{x}'_1,\mathbf{x}'_2}}{S_{\mathbf{x}'_1,\mathbf{x}'_2,\mathbf{x}'_3}} \Rightarrow S_{\mathbf{x}'_0,\mathbf{x}'_1,\mathbf{x}'_2} - \kappa S_{\mathbf{x}'_1,\mathbf{x}'_2,\mathbf{x}'_3} \approx 0, \tag{5.5}$$

where κ is the value of invariant ratio of areas. This imposes one constraint for the solution of $\mathbf{x}'_0$ as a quadratic equation. Similarly, all other possible pairs of triangles, with $\mathbf{x}'_0$ as the common vertex, can be selected to apply more constraints on $\mathbf{x}'_0$. Two such pairs of triangles are shown in Figure 5.2. Since there are 13 landmarks, there 66 possible triangle pairs. Thus, we have an over constrained system of quadratic equations to solve for the only unknown $\mathbf{x}'_0$

$$S_{\mathbf{x}'_0,\mathbf{x}'_i,\mathbf{x}'_j} - \frac{S_{\mathbf{x}_0,\mathbf{x}_i,\mathbf{x}_j}}{S_{\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k}} S_{\mathbf{x}'_i,\mathbf{x}'_j,\mathbf{x}'_k} \approx 0, \tag{5.6}$$

where $k, i, j = 1, \ldots, 13$ and $k \neq i \neq j$. This system of quadratic equations is solved using nonlinear least squares. Note that the unknown location is recovered independently of the other landmarks and does not rely on the computed locations in the previous frames, so the propagation of error is thus avoided. Therefore, given an initial correspondence among landmarks in the first frames of a model and test video, along with the known locations of landmarks in each frame of a model video, the unknown locations of the corresponding landmarks in the corresponding frames of the test video can be predicted.

5.2.2 Epipolar Constraint

Using the initial estimate of joint locations after the application of the affine constraint, we refine the predictions using the fundamental matrix between the two actors. Given the correspondences among the body joints in the first frames of the model and test videos, the fundamental matrix $\mathcal{F}$ is given by the relationship $\mathbf{x}_k \mathcal{F} \mathbf{x}'_k = 0$. Corresponding to a known location in the model video, the epipolar line constrains the search space of the joint location in the test video to the epipolar line 5.2. In [GS06], the epipolar line constraint was applied to minimize a search space of a joint.

For a given joint location in a particular frame of the model video, the fundamental matrix provides an epipolar line in the corresponding frame of a test video. We project the initial estimated location, from an affine constraint, onto this epipolar line to determine the final prediction for that point. This process is repeated for all the 13 points for every frame of the test video. The combination of both constraints provides the invariance to the variations in the viewpoint and the anthropometry of individuals in the two videos. It has been observed that the proposed constraints can handle substantial variation in the anthropometry, however, very large variations in the viewpoint can induce more error in the predictions. The appendix discusses details of a synthetic experiment performed for analysis of the viewpoint variance and the affine constraint.

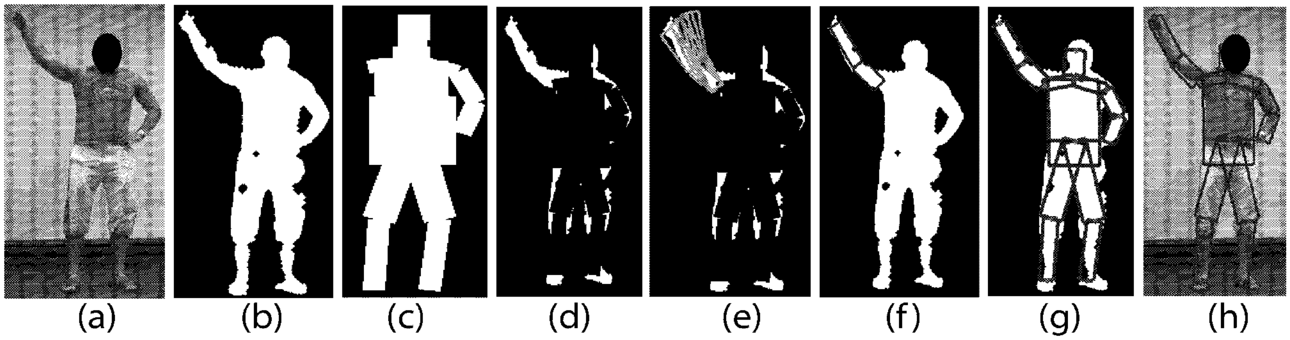


Figure 5.5: The steps for detecting right arm in the current frame are presented. (a) Input frame of the test video, (b) foreground silhouette extracted after the background subtraction, (c) template from the previous frame for detection of the right arm, (d) template mask applied to the silhouette, (e) a subset of the prediction positions used to search for the best fit for the arm, (f) selected best overlap between transformed prediction locations and the foreground silhouette, (g) other parts of the model are detected in a similar manner, and (f) the final detection results are superimposed on the input frame.

5.3 Joint Tracking

For tracking the exact locations of the 13 landmarks we use the predicted locations along with the features from the test video. The variations in the execution rate and the way in which action is performed is captured in this section by using the features observed in test video. There can be several options for extracting features from the test video, including interest points, color, texture, gradient, etc., but it has been observed that in the case of the human body, these features alone can be misleading. These features are greatly dependent upon illumination, resolution, type of clothing, etc. Instead, we use the foreground silhouettes obtained from background subtraction. We have performed successful experiments on sequences with no discriminative appearance features on the sequence as shown in Figure 5.11. Stick figures were used in the previous section, and now we use a cardboard model, as shown in the Figure 5.7(a), for tracking purposes. The connected landmark

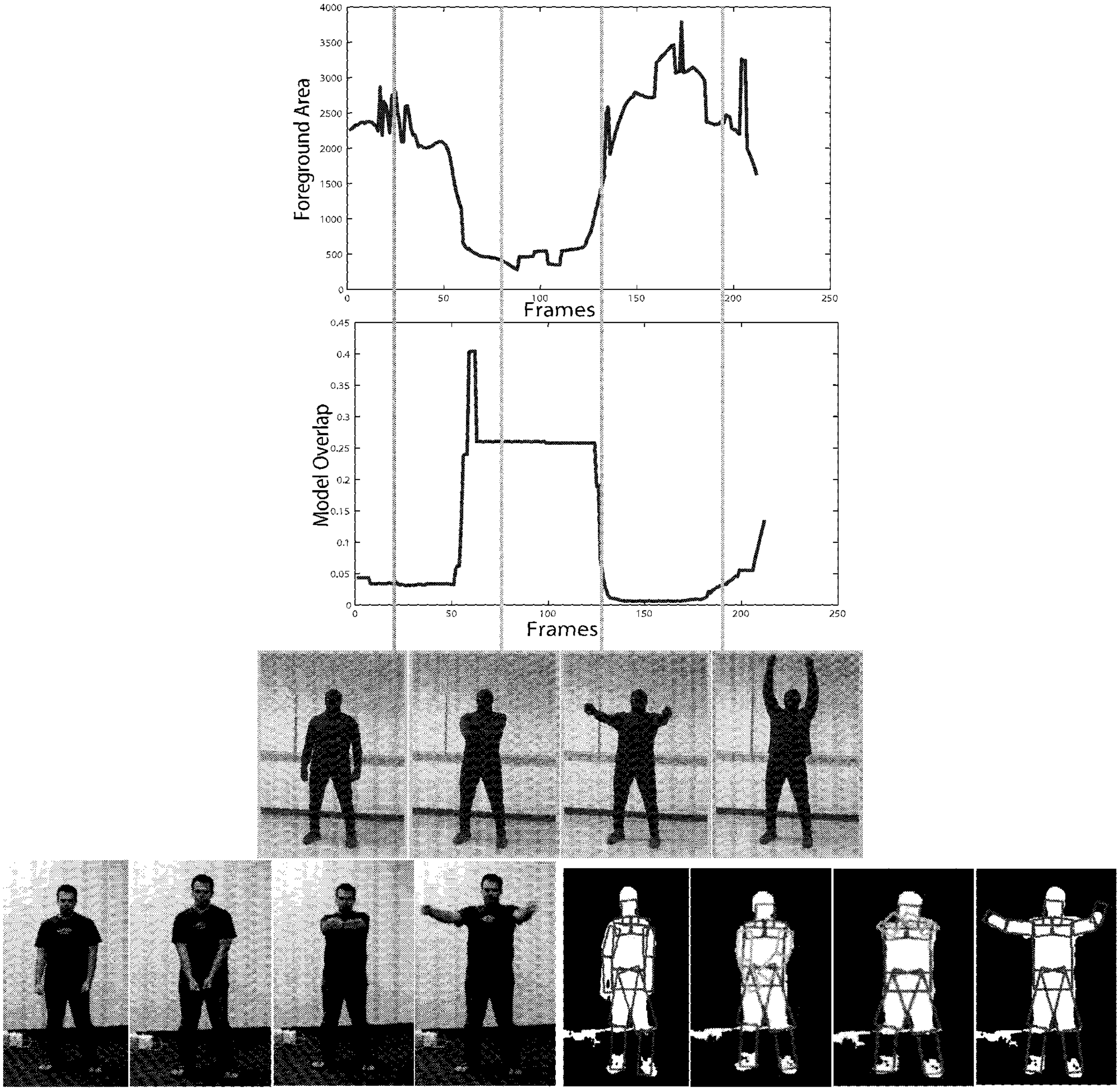


Figure 5.6: Occlusion Handling: (b)Video sequence at the bottom shows two arms coming in front of torso. Top row shows the plot for (a) plot of α_j^t: area of the foreground blob of the left arm and (b)plot of β_j^t: amount of overlap for model segment of left arm . Start and end of occlusion is detecting by the amount of change in these two measures. (d) Foreground silhouettes with tracking output during occlusion. The occluded locations are estimated from the model.

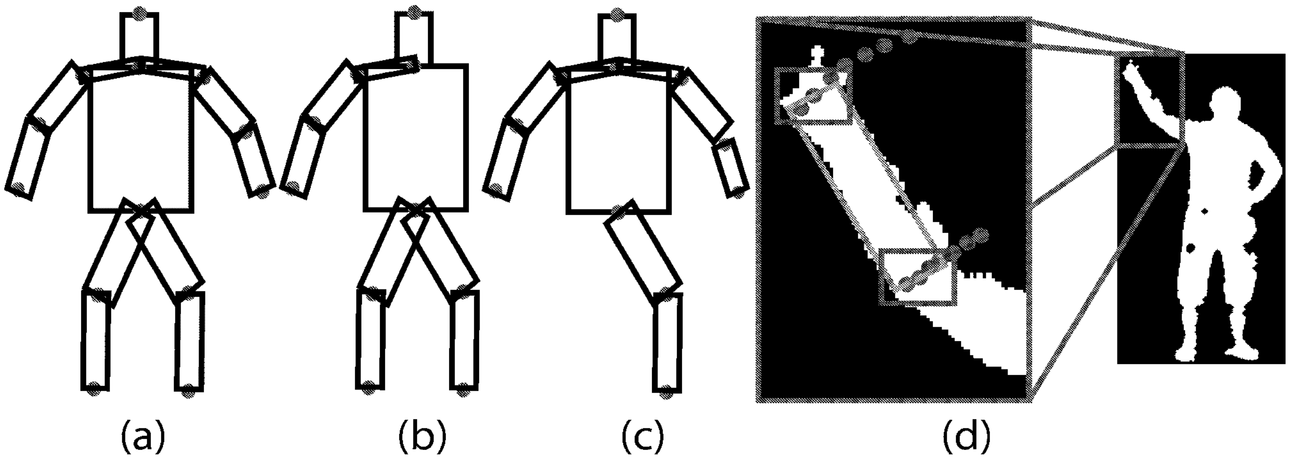


Figure 5.7: The templates for the cardboard model used for human body. (a) Complete template with 13 points and 12 segments. Templates are used to isolate the sections individual parts from the silhouettes. Two such examples are shown here for the detection of corresponding parts, (b) template for detecting left arm, (c) template for detecting right leg. (d) Blown up view of the arm shown on the left. Predicted points (red dots) are used for initialization in detection of the segment (green box) between two connected landmark points. The segment with optimal overlap with the foreground silhouette is used to find the best pair of predicted points within a temporal window. Spatial refinement follows with optimizing the overlap by varying length and rotation a small spatial window (blue box).

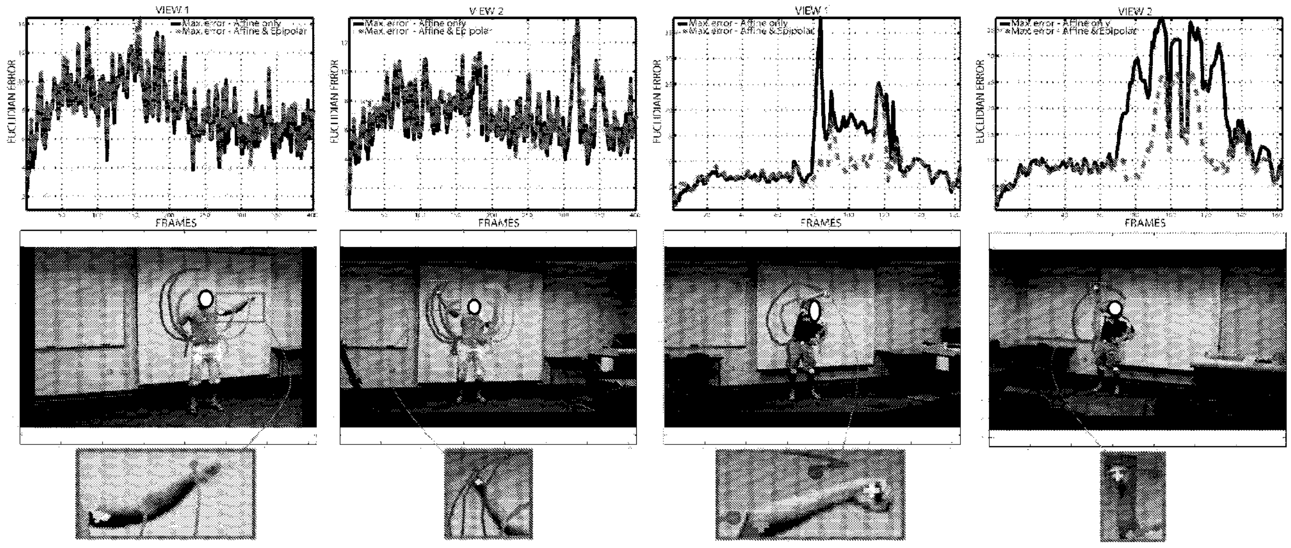
pairs in the stick figure are connected by a rectangle in this case. Figure 5.5(h) also shows a sample frame with the cardboard model superimposed on the original video. The segment between chest and belly point extends further down to cover the hip area. This extension is fixed throughout and is equal to the width of the torso. Our assumption is that the length of these segments can vary, but the width remains. For tracking a given segment in the test video, the best fit of a segment is determined by optimizing the overlap between the rectangle and its corresponding section of the foreground silhouette.

The cardboard model is initialized in the first frame using the available locations of the 13 landmarks. The width of each segment observed in the first frame is used for the rest of the video clip. For the following frames, the detection of each segment in the model is performed in the following order: torso (belly, chest, both shoulders), head, legs, and arms. The cardboard model is updated after detections from every frame. In a given frame we use templates derived from our cardboard model to isolate the section of the foreground silhouette that corresponds to the particular body segment. Figure 5.7(b,c) show sample templates for the detection of a left arm and right leg. After this region has been isolated, the next step is to fit the bounding box using the predicted joint locations from that body part. We use a set of predictions for every joint over a temporal window starting at the state from the previous frame. In our experiments we have used temporal windows ranging from 5 to 30 frames. The length of the temporal window can be an issue in case of joints which remain stationary for a long time and then move. We also handle this case by selecting the latest prediction giving the best detection.

The main steps involved in tracking the right arm from one of the experiments is shown in Figure 5.5. Steps are the same for tracking the other parts of the body. The cardboard model is updated every frame and the templates from the last frame are used for detection of all the parts in the current frame. This helps with detections in a case where the silhouette is completely closed and the predictions for one part cover the other part. Such templates are also helpful for detecting the area of every body part for occlusion handling. The isolated section of the silhouette is then used to find the exact location of the body part in the current frame as shown in the Figure 5.5(e). The details of the steps involved in this detection (based on the predictions in a temporal window) are illustrated in Figure 5.7(d). These detections are performed in the order starting from the inner body towards the limbs as mentioned earlier. Hence, for the arm, the shoulder is first detected with torso, followed by the elbow and the hand detected together. All the combinations of the predicted joint locations are tested for the best overlap between the foreground silhouette and the segment. This overlap is optimized by maximizing the foreground area while keeping the size of the segment controlled. The current predictions that provide the best segment fit are chosen for current joint. The location of this joint is then further improved by searching in a spatial window where the length and orientation of the segment is varied to optimize the final detection of the segment. This operation tries to accommodate for variations between the way the model and test actions are performed.

5.3.1 Occlusion Handling

An important issue in human body joint tracking is self occlusion. Here, we present our approach to handle self occlusion between different body parts. Figure 5.6 shows one example where both arms come in front of the torso and can not be observed in the foreground silhouette until they get out of occlusion near the shoulders as shown in the third frame. We use two measures for the detection of the beginning and ending of body part occlusion. The first measure is α_j^t which represents the area of the foreground blob corresponding to the segment j in frame t. A foreground blob is detected as shown in Figure 5.5(d) and the area of the corresponding blob is used for this measure. Secondly, the measure β_j^t represents the percentage of a detected segment occluded by other segments in the cardboard model (torso in this case). These parameter values are stored for every segment of the cardboard model. The values from these measures are drawn over a temporal



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

Figure 5.8: Quantitative and qualitative results of prediction phase. First row shows the maximum euclidian error in predictions. Using ground truth trajectories in view 1-(a), the trajectories in view 2-(b) were recovered, and vice versa. Similarly, trajectories in view 1-2(c-d) were recovered. Second row presents frames, 159, 319, 118, and 102 with the largest error in views 1-(a), 2-(b), 1-(c), and 2-(d), respectively. Third row shows the zoomed in fragments depicting the largest error.

window to overcome noise. Hence, we use the normalized accumulated change over time as shown in the condition for occlusion here

$$\frac{\sum_{i=t-\tau}^{t-1} \alpha_j^{i+1} - \alpha_j^i}{\alpha_j^{t-\tau}} < Th, \frac{\sum_{i=t-\tau}^{t-1} \beta_j^{i+1} - \beta_j^i}{\beta_j^{t-\tau}} > Th,$$

where Th is the percentage threshold (Th=0.7 in our experiments), and τ is the length of the temporal window. A positive Th value is used as a condition for entering occlusion, while negative Th values can be used for exiting occlusion. Figure 5.6(c) shows progression of the video summarized by four frames and the two measures α_j^t and β_j^t for the left arm are shown Figure 5.6(a) and (b).

During occlusion the locations of the joints are approximated by the corresponding predicted locations. The rate of action in test video can be different than that in the predicted model. Hence, the entry and exit instances are determined both from the silhouette feature (α_j^t) and the set of predicted locations. When the condition for end of occlusion has been satisfied, then the exact difference in execution rate through occlusion is determined. This value is then used to determine the joint locations by linear interpolation of the predicted locations during this occlusion interval. Figure 5.6(d) shows the estimated location of arms on the foreground silhouette during occlusion. We would like to mention here that the appearance (color and texture) features can be helpful here for better detection, but could be limited if the appearance of arms is not very discriminative.

5.4 Experimental Results

Several experiments were performed to analyze the performance of the proposed approach on several video sequences. The videos contained articulated motion, self occlusion, change in viewpoints, and a variety of actors.

We present both qualitative and quantitative results of the prediction phase from different videos. In the first set of experiments we analyze the error, which is the Euclidian distance between the ground truth and the locations computed by the presented approach. In the first experiment, an action was captured by a pair of cameras with little perspective variation. In both videos, body joints were manually marked and used as the ground truth. Using the joint locations in video 1 and initialized landmarks in the first frame of the video 2, the locations of the joints in video 2 were

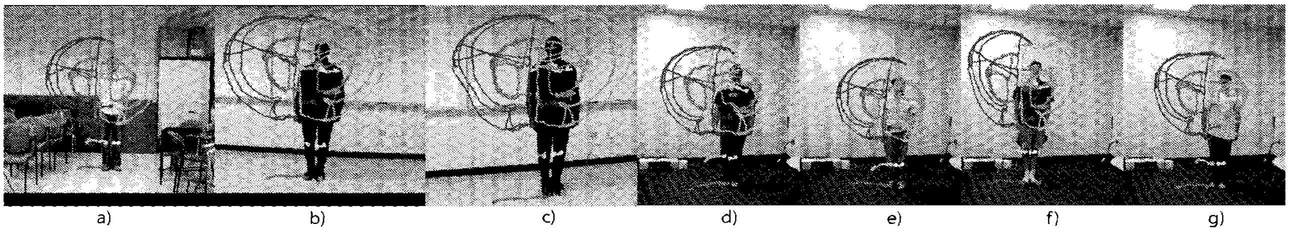


Figure 5.9: (a) Trajectories corresponding to the model action. Recovered trajectories are superimposed with the first frames of the video. (b) and (c) show the recovered trajectories in video captured by two wide-base line cameras. In (b)-view the perspective effect is less than in (c), so the trajectories are more similar to the model trajectories. (d-g) show the recovered trajectories corresponding to actors with significant change in the anthropometric measurements.

computed and compared to ground truth and vice versa. Predicted trajectories were computed in two different ways: using only an affine constraint and affine with epipolar constraint. First row of Figure 5.8 shows the maximum error in each frame. It can be noticed that under the first scene, the use of an epipolar constraint does not affect the error too much. Second row of Figure 5.8 shows selected frames with the largest error. Trajectories show locations computed in all the previous frames, while points represent the current locations and the ground truth locations are marked by crosses. For this scene, the largest error corresponds to the predictions of the left and right hands in frames 159 and 319 respectively. In the second experiment we repeated the same steps, but a new action was captured by a pair of cameras with wide baseline setting. The length of the video was 163 frames. It can be seen that in this case the epipolar constraint reduces the error significantly. The largest error corresponds to the predictions of the right hand in frames 118 and 102, respectively. The results of two experiments demonstrate the robustness to the view change. It also shows the relationship between the error of the prediction and perspectivity in the image.

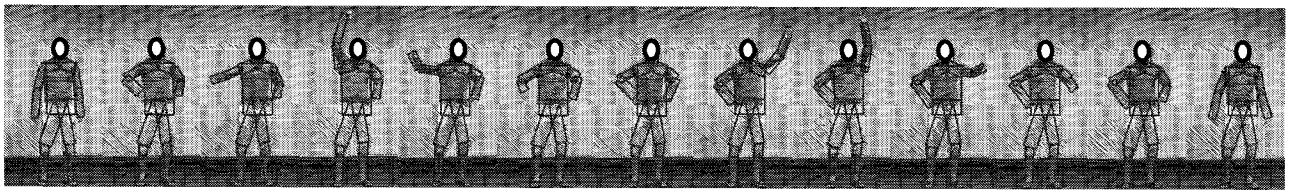


Figure 5.10: The output of tracking on 285 frames long sequence is shown here. The predicted trajectories are shown in Figure 5.8(b) The model video used for this action is shown in Figure 5.8(a). The tracking output was observed to be accurate.

Results of the tracking phase are presented for two different actions. The results on the first action are shown in Figure 5.10. This video contains 285 frames and the body parts were tracked correctly throughout all frames. The model view used for this test video is shown in Figure 5.8(a). The second action was more challenging as it contained higher variation in the viewpoint, anthropometry of individuals, and the execution rate. In addition to this, the action has more out of plane movement and self occlusion. Figure 5.9 shows the prediction results for this action and Figure 5.11 shows the tracking results.

5.5 Summary

A novel 2D model-based approach for computing human body joint motion has been presented. Its integration into human joint tracking has a variety of applications in the area of higher level action and activity analysis. The formulation of the geometric constraints on the geometry of human action is novel. Compared to the approaches that use either linear or non-linear filters for human motion modeling, the proposed approach is easier to adapt to any model, more robust to viewpoint changes, and does not require extensive training. Experimental results for the prediction

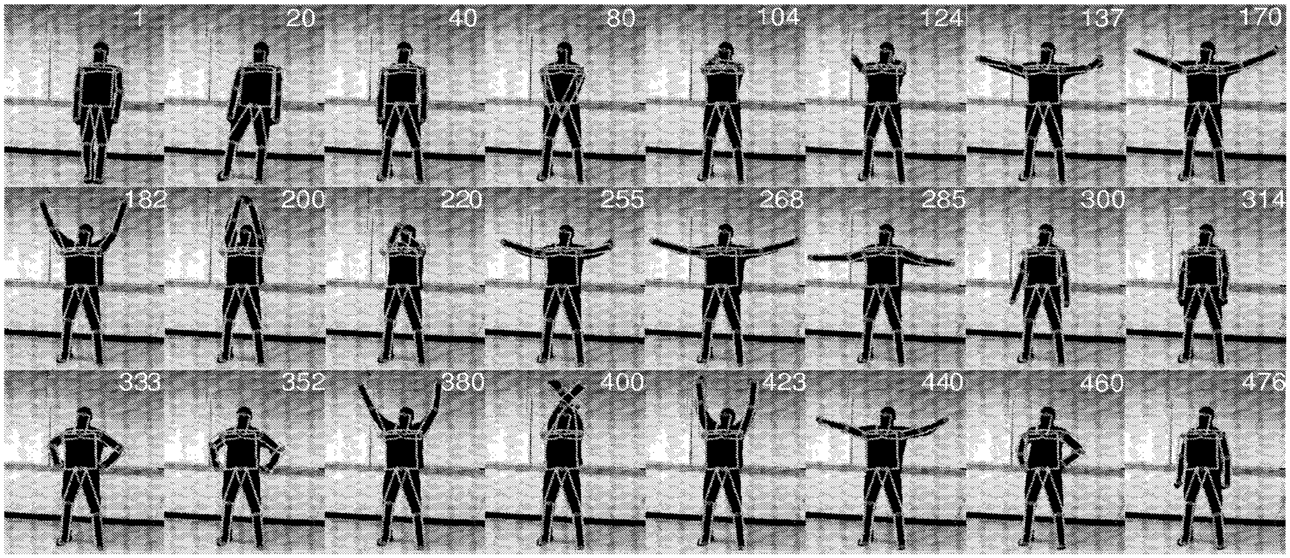


Figure 5.11: Test actor 1 (view 1). The final output of the tracking phase is presented. Model trajectories are shown in Figure 5.9(a). The test video is 476 frames long with articulated motion or arms and legs. There are two instances (frame 40-124, and 200-255) of full occlusion of both arms by the torso and one instance of partial occlusion between arms (frame 400). This video is also provided as the supplemental material.

and tracking phases demonstrated a great potential. The experiments support the thesis that the proposed approach can handle significant variations in the anthropometry and execution rate. At the same time, it was found to be quasi invariant to execution style and viewpoint variations.

CHAPTER 6
CONCLUSION AND FUTURE WORK

The objective was to find an approach for matching human actions that is both fully descriptive in terms of motion and is both invariant to view and execution rate. To our knowledge, this is the first work that expressly addresses the variability of human proportions. We make innovative use of epipolar geometry to propose a similarity measure between two sets of actions. Two related constraints were proposed and explored. Experimental validation of the proposed approach shows the versatility and importance of the results obtained in this work.

We have presented a spacetime projection model of cameras moving at constant velocities. In practice, the assumption of constant velocity is often reasonable for a short duration of time, especially when the camera is mounted on a vehicle such as aircraft, train, car, or on a robot. An important application for the ideas described in this paper is for prediction. When cameras move, the degree of overlap between their fields of view usually changes and when the fields of view become disjoint, the estimation of relative camera position becomes impossible. However, if the motion of the cameras follow some structured motion (like constant velocity), the ideas in this paper can be used to predict the fundamental matrix relating views even when their fields of view are disjoint. In this paper, we have investigated the relative geometry relating a pair of such cameras in planar and general scenes and defined the planar Galilean mapping and Galilean fundamental matrix that capture these relationships. Finally, we demonstrated how three known fundamental matrices are specializations of this matrix and could be readily recovered from the proposed fundamental matrix, and as a result, provide a unifying link between the classic fundamental matrix and the LP fundamental matrix. In addition, we described three new fundamental matrices that can also be recovered. In the future, we intend to investigate the application of the different motion models, such as a constant acceleration model, and study the relationships between three or more Galilean cameras.

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

We have proposed a novel 2D model-based approach for human body joint prediction and tracking which has a variety of applications in the area of higher level action and activity analysis. The formulation of the geometric constraints on the geometry of human action is novel. Compared to the approaches that use either linear or non-linear filters for human motion modeling, the proposed approach is easier to adapt to any model, more robust to viewpoint changes, and does not require extensive training. We have presented the results for prediction and tracking phases. The experiments support the thesis that the proposed approach can handle significant variations in the anthropometry and execution rate. At the same time it was found to be quasi invariant to execution style and viewpoint variations. The proposed framework can also be extended for an action recognition framework. In the future work, we plan to improve our tracking system. In particular, we want to extend the proposed tracking to actions involving more difficult cases of occlusion, find and integrate an optimized solution for propagation model states, and use a composite action representation to increase bank of action models.

LIST OF REFERENCES

[AC99] J. Aggarwal and Q. Cai. "Human Motion Analysis: A Review." In *Computer Vision and Image Understanding*, 1999.

[AH01] N. Johnson A. Galata and D. Hogg. "Learning variable length Markov models of behaviour." *Computer Vision and Image Understanding Journal*, **81**:398–413, 2001.

[Aki84] K. Akita. "Image sequence analysis of real world human motion." In *Pattern Recognition*, 1984.

[AP04] J. Aggarwal and S. Park. "Human Motion: Modeling and Recognition of Actions and Interactions." In *Second International Symposium on 3D Data Processing, Visualization and Transmission*, 2004.

[AS98] D. Ayers and M. Shah. "Recognizing human actions in a static room." In *Proc. IEEE Workshop on Applications of Computer Vision, WACV'98*, pp. 42–47, 1998.

[AS00] S. Avidan and A. Shashua. "Trajectory Triangulation: 3D Reconstruction of Moving Points from a Monocular Image Sequence." *PAMI*, 2000.

[AW03] S. Peleg A. Zomet, D. Feldman and A. Weinshall. "Mosaicing New Views: The Crossed-Slits Projection." *PAMI*, 2003.

[Bai82] Robert W. Bailey. *Human Performance Engineering: A Guide for System Designers.* Prentice-Hall, Inc., 1982.

[Bar03] A. Bartoli. "The Geometry of Dynamic Scenes - On Coplanar and Convergent Linear Motions Embedded in 3D Static Scenes." *Computer Vision and Image Understanding*, 2003.

[BD01] S. Blakemore and J. Decety. "From the perception of Action to the understanding of intention." In *Nature Reviews*, 2001.

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

[BGS05] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. "Actions as Space-Time Shapes." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1395–1402, 2005.

[BHB00] C. Bregler, A. Hertzmann, and H. Biermann. "Recovering non-rigid 3d shape from image streams." In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 13–15, 2000.

[BI98] A. F. Bobick and Y. Ivanov. "Action Recognition Using Probabilistic Parsing." In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 196–202, Santa Barbara, CA, 1998.

[BI05] O. Boiman and M. Irani. "Detecting Irregularities in Images and in Video." In *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2005.

[BJ98] M. Black and A. Jepson. "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation." pp. 63–84, 1998.

[BPW93] Norman Badler, Cary Philips, and Bonnie Webber. *Simulating Humans*. Oxford University Press, 1993.

[Bri82] Rizwan Bridger. *Human Performance Engineering: A Guide For System Designers*. Prentice-Hall, 1982.

[Bri95] Rizwan Bridger. *Introduction to Ergonomics*. McGraw-Hill, 1995.

[Bux03] H. Buxton. "Learning and Understanding Dynamic Scene Activity: A Review." In *Image and Vision Computing*, 2003.

[BWR92] J. Burns, R. Weiss, and E. Riseman. "The Non-Existence of General-Case View-Invariants." *Geometric Invariance in Computer Vision, Eds. J. Mundy and A. Zisserman*, 1992.

[BY95] M. Black and Y. Yacoob. "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion." In *IEEE International Conference on Computer Vision*, pp. 374 –381, June 1995.

[Cal00] J. Callahan. "The Geometry of Spacetime." *Springer-Verlag*, 2000.

[CBA96] L.W. Campbell, D.A. Becker, A. Azarbayejani, A.F. Bobick, and A. Pentland. "Invariant Features for 3D Gesture Recognition." In *Proceedings, International Conference on Automatic Face and Gesture Recognition*, pp. 157–162, 1996.

[CI00] Y. Caspi and M. Irani. "A step towards sequence-to-sequence alignment." In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 682–689, 2000.

[CS95] C. Cedras and M. Shah. "Motion-based Recognition: A Survey." In *Image and Vision Computing*, 1995.

[DB97] J. Davis and A. Bobick. "The representation and recognition of action using temporal templates." In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 928–934, 1997.

[DEP95] T.J. Darrell, I.A. Essa, and A.P. Pentland. "Task-specific Gesture Analysis in Real-Time using Interpolated Views." *IEEE Transactions on Pattern Analysis and Machine Vision*, 1995.

[DF02] T. Brodsky M. Trajkovic D. Weinshall, M-S. Lee and D. Feldman. "New view generation with a bi-centric camera." *ECCV*, 2002.

[DG99] J. Decety and J. Grezes. "Neural mechanisms subserving the perception of human actions." In *Trends in Cognitive Sciences*, 1999.

[DS94] J. Davis and M. Shah. "Three-dimensional gesture recognition." In *Proc. of Asilomar Conference on Signals, Systems, And computers*, 1994.

[DV99] A. Daems and K. Verfaillie. "Viewpoint-dependent Priming Effects in the Perception of Human Actions and Body Postures." In *Visual Cognition*, 1999.

[DW02] T. Pajdla D. Feldman and D. Weinshall. "On the Epipolar Geometry of the Crossed-Slits Projection." *ECCV*, 2002.

[DZ05] D.A. Forsyth D. Ramanan and A. Zisserman. *Strike A Pose: Tracking People by Findind Stylized Poses*. Proc. IEEE CVPR, 2005.

[EBM03] A.A. Efros, A.C. Berg, G. Mori, and J. Malik. "Recognizing action at a distance." *ICCV*, 2003.

[EKC82] Ronald Easterby, Kiren Kroemer, and Don Chaffin. *Anthropometry and Biomechanics - Theory and Application*. Plenum Press, New York, 1982.

[Far99] B. Farnell. "Moving Bodies, Acting Selves." In *Annual Review of Anthropology*, 1999.

[Fau92] O. Faugeras. "What can be seen in three dimensions with an uncalibrated stereo rig?" *Proceedings of the European Conference on Computer Vision*, 1992.

[FL01] O. Faugeras and Q.-T. Luong. "The Geometry of Multiple Images." *MIT Press*, 2001.

[Gav99] D. Gavrila. "The Visual Analysis of Human Movement: A Survey." *CVIU*, **73**(1):82–98, 1999.

[GH97] R. Gupta and R.I. Hartley. "Linear Pushbroom Cameras." *IEEE Transactions on Pattern Analysis and Machine Vision*, 1997.

[GK03] S. Baker G.K.M. Cheung and T. Kanade. "Shape-From-Silhouette of Articulated Objects and its Use for Human Body Kinematics Estimation and Motion Capture." *CVPR*, 2003.

[Gol70] A. Goldman. "A Theory of Human Action." In *Englewood Cliffs, Prentice Hall*, 1970.

[GS89] Kristine Gould and Mubarak Shah. "The Trajectory Primal Sketch: A Multi-Scale Scheme for Representing Motion Characteristics." In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 79–85, San Diego, June 1989.

[GS06] A. Gritai and M. Shah. *Tracking of Human Body Joints Using Anthropometry.* ICME, 2006.

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

[Har92] R. Hartley. "Estimation of Relative Camera Positions for Uncalibrated Cameras." *ECCV*, 1992.

[Her79] M. Herman. "Understanding body postures of human stick figures." In *PhD Thesis, University of Maryland*, 1979.

[HHD00] I. Haritaoglu, D. Harwood, and L. Davis. "W4: Real-time surveillance of people and their activities." **22**(8):809–830, 2000.

[HK00] M. Han and T. Kanade. "Reconstruction of a Scene with Multiple Linearly Moving Objects." *CVPR*, 2000.

[Hog84] D.C. Hogg. *Interpreting Images of a Known Moving Object.* PhD thesis, University of Sussex, 1984.

[HW88] B. Horn and E. Weldon. "Direct Methods for Recovering Motion." *International Journal of Computer Vision*, 1988.

[HZ00] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, September 2000.

[IC03] E. Hunter I.Mikic, M. Traverdi and P. Cosman. "Shape-From-Silhouette of Articulated Objects and its Use for Human Body Kinematics Estimation and Motion Capture." *CVPR*, 2003.

[JBY96] S. Ju, M. Black, and Y. Yacoob. "Cardboard people: A parameterized model of articulated image motion." In *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 38–44, 1996.

[Joh73] G. Johansson. "Visual Perception of Biological Motion and a Model for its Analysis." *Perception and Psychophysics*, **14**(2):210, 211 1973.

[JS94] W. Liao J. Aggarwal, Q. Cai and B. Sabata. "Articulated and Elastic Non-Rigid Motion: A Review." In *Workshop on Motion of Non-Rigid and Articulated Objects*, 1994.

[KKP03] A. Koschan, S. Kang, J. Paik, B. Abidi, and M. Abidi. "Color active shape models for tracking non-rigid objects." *Pattern Recognition Letters*, 2003.

[LBP05] I. Laptev, S. J. Belongie, P. Prez, and Josh Wills. "Periodic Motion Detection and Segmentation via Approximate Sequence Alignment." In *Proceedings of the IEEE International Conference on Computer Vision*, 2005.

[LG05] H. Li and M. Greenspan. "Multi-Scale Gesture Recognition from Time-Varying Contours." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 236–243, 2005.

[LSS05] B. Leibe, E. Seemann, and B. Schiele. "Pedestrian detection in crowded scenes." *CVPR*, 2005.

[LT03] W. Hu L. Wang and T. Tan. "Recent Development in Human Motion Analysis." In *Pattern Recognition*, 2003.

[MG01] T. Moeslund and E. Granumm. "A survey of computer vision based human motion capture." In *Computer Vision and Image Understanding*, 2001.

[MHK06] T.B. Moeslund, A. Hilton, and V. Krüger. "A survey of advances in vision-based human motion capture and analysis." *CVIU*, 2006.

[Mis66] L. Von Mises. "Human Action: A Treatise on Economics." In *Chicago: Henry Regnery*, 1966.

[MOB05] A. Micilotta, E. Ong, and R. Bowden. *Detection and tracking of humans by probabilistic body part assembly.* In Proc. British Machine Vision Conf., 2005.

[NA94] S. Niyogi and E.H. Adelson. "Analyzing and recognizing walking figures in XYT." In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 469–474, 1994.

[NB05] S. Venkatesh N. Nguyen, D. Phung and H. H. Bui. "Learning and detecting activities from movement trajectories using the hierarchical hidden Markov models." In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 2005.

[NOM98] A. Nishikawa, A. Ohnishi, and F. Miyazaki. "Description and recognition of human gestures based on the transition of curvature from motion images." In *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 552–557, 1998.

[OHG02] N. Oliver, E. Horvitz, and A. Garg. "Layered representations for human activity recognition." In *Fourth IEEE Int. Conf. on Multimodal Interfaces*, pp. 3–8, 2002.

[ORP99] N. Oliver, B. Rosario, and A. Pentland. "A bayesian computer vision system for modeling human interactions." In *Proceedings of ICVS99*, Gran Canaria, Spain, January 1999.

[PC02] V. Parameswaran and R. Chellappa. "Quasi-Invariants for Human Action Representation and Recognition." *International Conference on Pattern Recognition*, 2002.

[PC03] V. Parameswaran and R. Chellappa. "View Invariants for Human Action Recognition." *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2003.

[PCar] V. Parameswaran and R. Chellappa. "Using 2D Projective Invariance for Human Action Recognition." *International Journal of Computer Vision*, to appear.

[PF02] R. Plankers and P. Fua. "Model-based silhouette extraction for accurate people tracking." *ECCV*, 2002.

[PN94] R. Polana and R.C. Nelson. "Detecting activities." *Jl. of Visual Communication and Image Representation*, **5**:172–180, 1994.

[Pri97] W. Prinz. "Perception and action planning." In *European Journal of Cognitive Psychology*, 1997.

[Ras80] R. Rashid. "Towards a system for the interpretation of moving light display." In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1980.

[RB80] J. O' Rourke and N. Badler. "Model-Based Image Analysis of human motion Using Constraint Propagation." In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1980.

[RC82] K. Kroemer R. Easterby and D. Chaffin. "Anthropometry and Biomechanics - Theory and Appplication." In *Plenum Press*, 1982.

[RCK06] M. Roh, B. Christmas, J. Kittler, and S. Lee. "Robust player gesture spotting and recognition in low-resolution sports video." *ECCV*, 2006.

[RGS03] C. Rao, A. Gritai, M. Shah, and T. Syeda-Mahmood. "View-Invariant Alignment and Matching of Video Sequences." *Proceedings of the IEEE International Conference on Computer Vision*, pp. 939–945, 2003.

[RM87] H. Baker R. Bolles and D. Marimont. "Epipolar-plane Image Analysis: An Approach to Determining Structure From Motion." *International Journal of Computer Vision*, 1987.

[RMR04] T.J. Roberts, S.J. McKena, and I.W. Ricketts. *Human Pose Estimation using Learnt Probabilistic Region Similarities and Partial Configurations.* ECCV, 2004.

[RS01] C. Rao and M. Shah. "View Invariance in Action Recognition." In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii, Dec. 2001.

[RT02] C. Schmid R. Ronfard and B. Triggs. "Learning to Parse Pictures of People." *ECCV*, 2002.

[SC78] H. Sakoe and S. Chiba. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition." *IEEE Transactions on ASSPR, Vol. 26, No.1*, 1978.

[SD97] S. M. Seitz and C. R. Dyer. "View-invariant analysis of cyclic motion." *International Journal of Computer Vision*, **25**:1–25, 1997.

[SI05] E. Shechtman and M. Irani. "Space-Time Behavior Based Correlation." In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2005.

[Sid04] H. Sidenbladh. "Detecting human motion with support vector machines." *ICPR*, 2004.

[SN] Y. Singer S. Fine and N.Tishby. "The hierarchical hidden Markov model: Analysis and applications.".

[SP96] T. Starner and A. Pentland. *Motion-Based Recognition*, chapter Real-Time American Sign Language Recognition from Video Using Hidden Markov Models. Computational Imaging and Vision Series. Kluwer Academic Publishers, 1996.

[SS06] F. Rafi S. Khan and M. Shah. "Where was the Picture Taken: Image Localization in Route Panoramas using Epipolar Geometry." *ICME*, 2006.

[ST03] C. Sminchisescu and B. Triggs. *Estimating articulated human motion with covariance scaled sampling*. IJRR, 2003.

[Stu02] P. Sturm. "Structure and Motion for Dynamic Scenes - the case of Points Moving in Planes." *ECCV*, 2002.

[Ver92] K. Verfaillie. "Variant Points of View on Viewpoint Invariance." In *Canadian Journal of Psychology*, 1992.

[VR96] L. Fogassi V. Gallese, L. Fadiga and G. Rizzolatti. "Action recognition in the premotor cortex." In *Brain*, 1996.

[WN05] B. Wu and R. Nevatia. "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detection." *ICCV*, 2005.

[WY06] Y. Wu and T. Yu. "A field model for human detection and tracking." *PAMI,vol.28(5)*, 2006.

[WZ02] L. Wolf and A. Zomet. "Sequence to Sequence Self-Calibration." In *Proceedings of the European Conference on Computer Vision(ECCV)*, Copenhagen, May 2002.

[XM05] A. C. Berg X. Ren and J. Malik. "Recovering Human Body Configuration using Pairwise Constraints between Parts." *ICCV*, 2005.

[YA98] M. Yang and N. Ahuja. "Extracting gestural motion trajectories." In *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 10–15, 1998.

[YH05] R. Sukthankar Y. Ke and M. Hebert. "Efficient Visual Event Detection using Volumetric Features." In *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2005.

[YOI95] J. Yamato, J. Ohya, and L. Ishii. "Recognizing human action in time-sequential images using hidden markov model." In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 624–630, 1995.

[YS05a] A. Yilmaz and M. Shah. "Actions As Objects: A Novel Action Representation." *IEEE Proceedings on the Interntional Conference on Computer Vision and Pattern Recognition*, 2005.

[YS05b] A. Yilmaz and M. Shah. "Actions Sketch: A Novel Action Representation." *CVPR*, 2005.

[YS05c] A. Yilmaz and M. Shah. "Recognizing human actions in videos acquired by uncalibrated moving cameras." *ICCV*, 2005.

[YXC97] J. Yang, Y. Xu, and C.S. Chen. "Human action learning via Hidden Markov Model." *IEEE Trans. on System, Man, and Cybernetics*, **27**(1):34–44, 1997.

[Zat02] V. Zatsiorsky. "Kinematics of Human Motion." In *Human Kinetics*, 2002.

[Zha98] Z. Zhang. "Determining the Epipolar Geometry and its Uncertainty: A Review." *IJCV*, 1998.

[ZI01] L. Zelnik-Manor and M. Irani. "Event-Based Analysis of Video." In *IEEE Conference on Computer Vision and Pattern Recognition*, Dec. 2001.