ABSTRACT

Title of dissertation: HUMAN MOVEMENT ANALYSIS: BALLISTIC DYNAMICS, AND EDGE CONTINUITY FOR POSE ESTIMATION

Shiv Naga Prasad Vitaladevuni
Doctor of Philosophy, 2007

Dissertation directed by: Professor Larry S. Davis
Department of Computer Science

We present two contributions to human movement analysis: (a) a ballistic dynamical model for recognizing movements, and (b) a model for coupling edge continuity with contour matching.

We describe a Bayesian approach for visual analysis of ballistic hand movements, namely reaches and strikes. These movements are most commonly used for interacting with objects and the environment. One of the key challenges to recognizing them is the variability of the target-location of the hand - people can reach above their heads, for something on the floor, etc. Our approach recognizes them independent of the movement's target-location and direction by modelling the ballistic dynamics. A video sequence is automatically segmented into ballistic subsequences without tracking the hands. The segments are then classified into strike and reach movements based on low-level motion features. Each ballistic segment is further analyzed to compute qualitative labels for the movement's target-location and direction. Tests are presented with a set of reach and strike movement sequences.

We present an approach for whole-body pose contour matching. Contour matching in natural images in the absence of foreground-background segmentation is difficult. Usually an asymmetric approach is adopted, where a contour is said to match well if it aligns with a subset of the image's gradients. This leads to problems as the contour can match with a portion of an object's outline and ignore the remainder. We present a model for using edge-continuity to address this issue. Pairs of edge elements in the image are linked with affinities if they are likely to belong to the same object. A contour that matches with a set of image gradients is constrained to also match with other gradients having high affinities with the chosen ones. A Markov Random Field framework is employed to couple edge continuity and contour matching into a joint optimization process. The approach is illustrated with applications to pose estimation and human detection.

HUMAN MOVEMENT ANALYSIS: BALLISTIC DYNAMICS, AND EDGE CONTINUITY FOR POSE ESTIMATION

by

Shiv Naga Prasad Vitaladevuni

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Professor Larry S. Davis, Chair/Advisor
Professor Yiannis Aloimonos
Professor David Jacobs
Professor Ramani Duraiswami
Professor Benjamin Kedem

ACKNOWLEDGMENTS

I would like to acknowledge my advisor, friends and colleagues for making graduate life exciting and rewarding. Prof. Larry Davis gave me his support and trust through the ups and downs of Ph.D. work. His door was always open for talking about new ideas, and he presented me with a challenging research problem every six months. Prof. Davis provided ample opportunities to collaborate with other graduate students. I hope to carry forward his emphasis on personal integrity and pragmatism throughout my life. I would also like to thank Prof. David Jacobs and Prof. Ramani Duraiswami for their encouragement and the many interesting discussions. Prof. Jacobs' course on image segmentation started off my interest in perceptual organization which developed into the work on edge continuity. I found Prof. Duraiswami to be an accessible reference for my questions in mathematics. Both Prof. Jacobs and Prof. Duraiswami highlighted the need for qualitative understanding of concepts and encouraged creative and venturesome thinking. Prof. Yiannis Aloimonos expressed his interest in my work on human movement analysis early on, and gave many useful insights during his course. Dr. Yaser Yacoob gave me the psychology book on human movements that became the basis for the ballistic dynamics model.

The work on ballistic movements was done in collaboration with Vili Kellokumpu. He also made my stay at the University of Oulu memorable. Mohamed

Hussein helped with the human detection experiments. I would like to acknowledge Dr. Vinay Shet, Gaurav Aggarwal, Gopi Suvanam, Ashok Veeraraghavan, Narayanan Ramanathan, Aniruddha Kembhavi, Abhinav Gupta, Vlad Morariu, Behjat Siddiqui and Son Dinh Tran for making my stay at UMD a fun experience.

I dedicate this thesis to my family. Lavanya gave me her unflinching love and positivity during the trials and tribulations of graduate life. My parents, Dr. V. R. V. Ramanan and Dr. Kastala Jayasri, *Didi* and *Jiju* provided love and trust, which made everything possible.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Automated visual recognition of human movements is a principal enabling technology for video-based activity analysis and human computer interaction systems. The applications are wide-ranging, from medical diagnosis and monitoring the well-being of the aged, to multimedia content analysis and surveillance systems. We present contributions to pose-estimation and dynamical modelling for movement recognition systems.

Human pose-estimation addresses the problem of identifying the pose of humans in images. For example, a traffic hand-signal recognition system would estimate the position of the arms of the person directing traffic to recognize gestures such as "turn left" and "turn right". A popular approach for estimating the pose is to collect example silhouettes of humans in different poses, and compare them with the test image. Typically, the body's outline is represented with a contour which is compared with the image edges. However, edge clutter present in natural images complicates this task. We explore the utility of edge continuity for improving the estimation accuracy. In doing so, we build upon more than three decades of research in perceptual organization and edge continuity. An edge affinity model is presented that combines edge continuity with color statistics. This is combined with an extension to the Chamfer matching approach into a unified pose-estimation algorithm.

We show the efficacy of the model by applying it to human pose estimation and as part of a gesture recognition system. This concept is further developed to couple edge affinity and contour matching in a joint optimization problem using Markov Random Fields (MRFs). It is illustrated with a human detection task.

Tracking human poses is one of the principal challenges in movement recognition [53]. The reasons include ambiguity in pose estimation due to noisy edges and pose singularities, and errors in the dynamics. A number of recent studies have addressed this issue by relying on low-level image and motion features that avoid tracking. The emphasis is on applying machine learning techniques to model the statistics of these features. However, most such approaches perform recognition on relatively distinct action classes such as kneeling, sitting, standing, kicking, etc. We present an approach that attempts to take a middle-ground between explicit pose-tracking and employing solely low-level features. Psychological studies of human movements propose that common human movements are ballistic in nature. When humans become adept at executing an action, the movement speed increases, resulting in impulsive propulsion. This, in turn, results in a simplified trajectory of the hand and other body parts. The high movement speeds that make pose-tracking hard also provide characteristic signatures to the motion features that can be modelled with machine learning schemes. We develop a Bayesian model for ballistic movements to perform recognition without pose-tracking. Continuous videos are automatically segmented into individual ballistic movements, each of which are recognized based upon the dynamics of low-level cues, and the starting and ending pose of the person.

Chapter 2 reviews psycho-kinesiological studies on ballistic movements and describes the key observations used in our work. One of its secondary aims is to present the findings to the computer vision community from an action recognition perspective for future research. Chapter 3 presents the Bayesian model for ballistic movements and illustrates it with experiments on motion capture data. The probabilistic framework provides robustness and allows the approach to be potentially combined with parallel research in movement recognition. These ideas are further developed into a video-based movement recognition system, described in Chapter 4. The edge affinity model for pose matching is presented in Chapter 5, and the MRF-based framework is described in Chapter 6.

Chapter 2

Ballistic Hand Movements

The objective of visual analysis of human activity is to automatically understand the intentions guiding human actions observed in video and to identify stylistic attributes. It has numerous applications, such as analyzing customer behavior in retail stores, monitoring the well-being of senior citizens, assigning semantic labels to videos, and surveillance. Interactions with objects and the environment forms a key component of human activities. Consider an everyday scenario, such as a person boiling water for tea in a kitchen. This activity may be considered to consist of a sequence of actions such as opening a cupboard, reaching for the pot kept inside, putting the pot on a burner and so on. Analyzing a video recording of the activity of "brewing tea" would involve recognizing these individual movements - this is the focus of our study. This chapter presents motivations and challenges to the automatic recognition of human movements, and the limitations of state-of-the-art vision approaches. It then describes certain observations reported in psycho-kinesiological studies of human movements, and how these can be used as leverage for automatic recognition. These ideas are further developed in Chapters 3 and 4 into a video-based movement recognition system.

Continuing with our illustrative activity of brewing tea, a typical adult who is familiar with the kitchen's layout would execute the actions efficiently, with rapid

and coordinated body movements. In terms of dynamics, such movements have two characteristics:

1. They involve impulsive propulsion with rapid acceleration and deceleration [79, 51].

2. Human adults are capable of accurately (and unconsciously) planning the execution of reach movements before the commencement of motion. A large majority of such movements are completed with little or no mid-course correction. For instance, a number of models proposed in the psychology literature hypothesize that the dynamics of the hand remain fixed for the course of the movement, e.g. [28, 85, 80].

Due to their impulsive nature, these movements are referred to as "Ballistic" in psycho-kinesiology. Ballistic movements form a large portion of human interactive actions, evidenced by the extensive studies in psychology e.g., [79, 51, 28, 85, 37, 80, 60, 35, 27, 6, 21]. These movements include:

(a) *Reach actions:* e.g., reach-to-grasp, pointing gestures, placing objects.

(b) *Strike actions:* e.g., punching and throwing.

A system capable of recognizing individual reaches and strikes would enable the analysis of activities as a sequence of such movements. This forms the principal motivation of our study. For the design objectives, the following constraints are imposed on the system:

1. Use single camera video data. Do not assume the availability of the body's pose information.

2. The movements should be recognized independent of the hand's target location. E.g., the person could reach for something on the floor, above the head, to the left, etc. All of these instances should be recognized as reach movements and then additional labels must be computed to describe their target's location. Similarly, the strike movements must be recognized irrespective of where and in which direction the person punches or throws.

3. The movements may be executed as part of a continuous activity.

4. The person's pose with respect to the camera may vary between different instances of the movements.

These constraints are illustrated in the movement sequence shown in Figure 2.1 in which a person picks up an object from the floor and places it at another location on the floor. This action consists of 4 movements: bend down to grasp the object, pick it up, step to the other location and bend to place the object on the floor. The movements have different targets but have reach-dynamics as the common denominator. Predictably, change in the movement's target results in change in the trajectory followed by the hand and other body-parts. The two bend-and-reach movements have different body orientations w.r.t. the camera, resulting in variations in the poses' appearance. Figure 2.1 also shows the labels computed by our system.

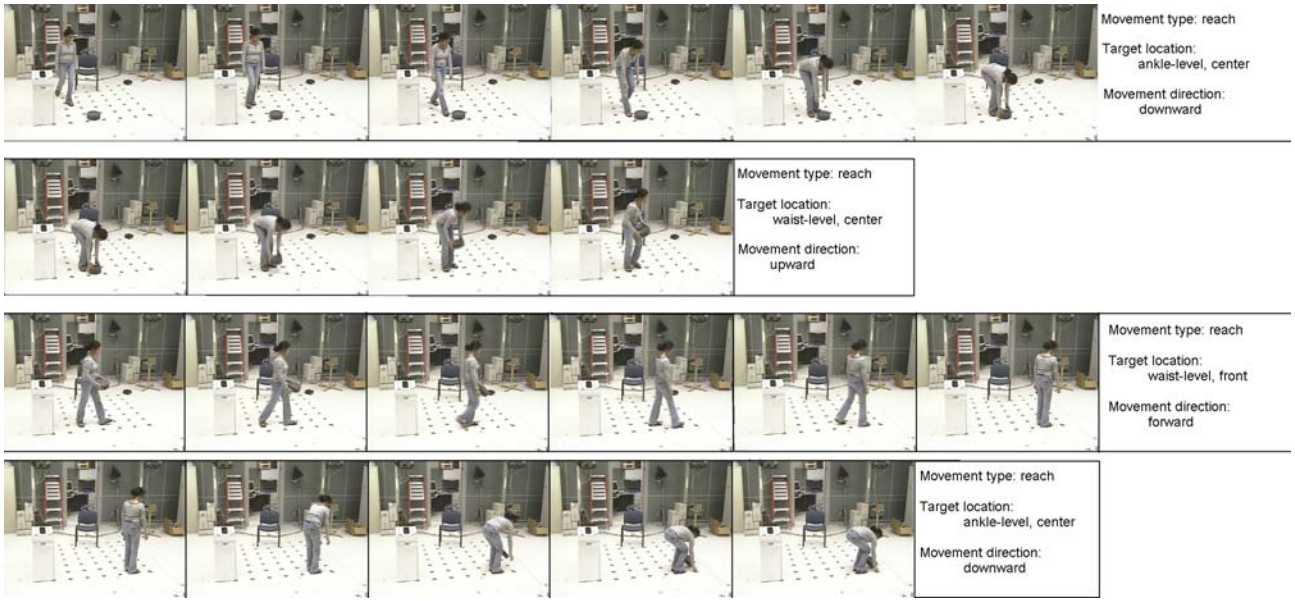


Figure 2.1: Temporal segmentation and labels generated for a movement sequence of 52 frames. To save space, only every third frame is shown. (Best viewed in color.)

2.1 Current Approaches to Visual Recognition of Human Movements

We identify two broad categories of human movement recognition approaches:

- Approaches that either track the body-parts, maintain a state for the poses, or assume availability of body trajectories. These include control-theoretic systems, such as those based on Hidden Markov Models (HMMs) [64], Switching Linear Dynamical Systems (SLDSs) [68]. There are also algebraic approaches that analyze segments of body-part trajectories. For example, Sheikh et al. model actions as sub-spaces of body trajectories [77].

- Recently, a number of studies have addressed action recognition by modelling the statistics of low-level motion features, e.g., [94, 89, 76]. These approaches do not track individual body parts.

Human movement analysis by tracking body poses may be viewed as a process of iterating over two steps:

1. Estimate the pose at time t based on observations from the video frame.

2. Predict the pose at time $t + 1$ using the current pose estimate and a model of the dynamics. This prediction is combined with the observations from the video frame at time $t + 1$ to estimate the pose at $t + 1$.

Precise pose estimation would enable accurate computation of the movement dynamics, enabling correct recognition. However, human pose estimation has proven to be one of the hardest problems in vision, and remains the principal stumbling block in movement analysis [31, 53]. Reasons for the problem's complexity include the large

number of degrees of freedom of the human body, singularities in pose-appearance from single-camera view and edge clutter from clothing. A popular approach to address this issue is to formulate movements as stochastic processes and perform probabilistic recognition. For example, in Hidden Markov Model (HMM) based approaches [64], poses are considered to be the states of a hidden random variable which is observed stochastically through image features. The dynamics constitute the transitions over these states. Here, prior information about the dynamics reduces uncertainty in pose estimation. HMMs and their variants have been shown to be effective for gesture recognition [91, 24], gait analysis [68], etc.

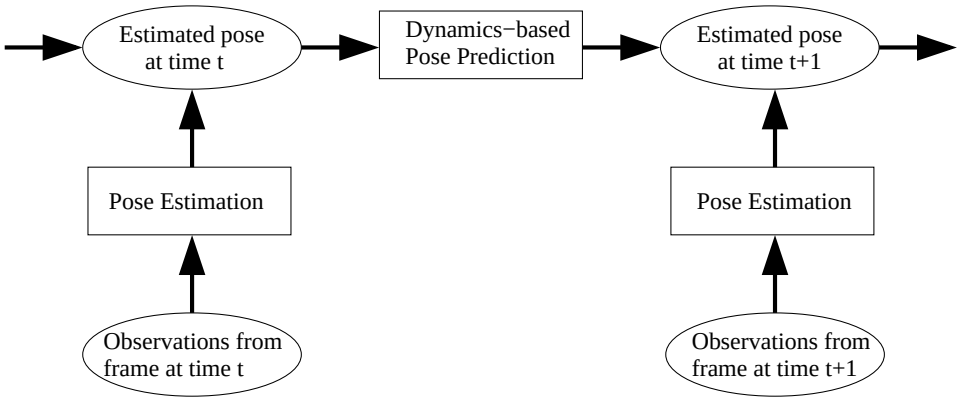


Figure 2.2: Schematic of the movement analysis process. Observations made from the video frame at time t are analyzed to estimate the pose at t. The dynamical model of the movement, and the pose observations at $t+1$ determine the pose at time $t+1$.

Sub-space methods model actions by computing algebraic invariants with respect to variation in camera viewpoint, style and speed. Suppose $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ were the hand's trajectories during two instances of some action. Then the action is modelled as the sub-space, F, such that $F\mathbf{x}_1(t) = F\mathbf{x}_2(t)$. This has been employed for computing view-invariants [95], as well as style-invariants [77].

Several studies have sought to bypass explicit pose estimation by directly modelling actions using low-level image intensity and motion features. Here, the emphasis is on applying machine learning techniques to learn the statistics of the features for the actions of interest. These techniques have been shown to be robust to image noise, small variations in illumination, view and movement style. Shechtman and Irani propose to model behavior with statistics of image intensity gradients in spatio-temporal volumes [76]. Yilmaz et al. extract contours of humans and construct 3D volumetric shapes by stacking them over time [94]. The shapes' forms model the underlying actions. Related to this, Bobick and Davis construct temporal templates of silhouettes and match them for recognition [11]. Weinland et al. employ 3D reconstructions of the body in a similar manner [89]. The features employed in these studies are highly dependent upon the viewpoint and trajectory of the movement. For instance, it is not clear if they would be able to generalize between reach movements towards different targets - reaching for something on the floor versus at shoulder level.

To summarize, approaches based on dynamical models and pose-tracking suffer from the ambiguities in pose-estimation. However, the dynamical models provide the ability to generalize over variations in viewpoint and movement targets. In contrast, approaches relying on low-level intensity and motion features are robust as they do not require pose estimation. However, it is not clear if they can generalize over movement targets.

We propose to exploit the ballistic nature of reach and strike movements to recognize them. The poses are not tracked explicitly. Instead, low-level motion

features are employed to represent the movement dynamics. Pose-estimation is performed at the start and end of the movement to compute labels. This can be viewed as a combination of pose and low-level feature analysis.

The next section reviews observations made in psycho-kinesiological studies regarding the following questions:

1. What are the body-parts that should be tracked to analyze reach and strike movements? To what level of detail must the poses be estimated? Do the joint-angles of the arms have to be estimated? A system requiring only a coarse pose-estimate would be more robust and practical.

2. What is the reference frame in which the dynamics should be analyzed? E.g., should the reference be body-centric or world-centric?

3. What structure of the dynamical model is suitable for analyzing reach and strike movements?

2.2 Psycho-kinesiological Studies of Ballistic Movements

Psychologists have proposed two models for limb propulsion [79]: ballistic movements and mass-spring movements, which form two ends of a spectrum of human movements. Ballistic movements involve impulsive propulsion of the limbs. There is an initial impulse accelerating the hand/foot towards the target, followed by a decelerating impulse to stop the movement. There is no mid-course correction. Reaching, striking and kicking are characteristically ballistic movements [79, 51]. In the mass-spring model, the limb is modelled as a mass connected to springs

(the muscles). The actuating force is applied over a period of time rather than impulsively [79, 8]. Steady pushing, pulling, and many communicative gestures fall into this category.

Rapid, practiced movements usually follow the ballistic model - the majority of the movements observed in everyday activity are ballistic. Slower, smoother movements are modelled well as a mass-spring system [43]. When a subject becomes confident about movements, the speed is usually high. High speeds tend to have impulsive propulsion, making mid-course corrections difficult.

For movements following the mass-spring model, the limb is in dynamic equilibrium during the movement. Therefore, the trajectories can be altered at any time - enabling them to be more complex than ballistic movements.

There are two differences between ballistic and mass-spring models of movements that are relevant for recognizing human actions:

1. Ballistic movements have a simpler structure. Often, the starting and ending positions of the limbs are sufficient to specify the trajectory of a ballistic movement. In contrast, the mass-spring model allows for complicated trajectories. For example, drawing a figure '8' with the hand, moving the hand in a circle to signal "start engine", etc.

2. Reaching, striking, waving, kicking, etc., which are predominantly ballistic, are common actions encountered during surveillance. These have highly variable target locations. Mass-spring movements, especially communicative gestures, have higher spatial consistency.

Due to the relatively simple structure of force actuation - acceleration followed by deceleration - ballistic movements have a characteristic "bell"-shaped velocity profile [79]. Figure 2.3 shows velocity profiles of some mass-spring and ballistic hand movements. Plots of different movement instances are shown in different colors for discernibility. The mass-spring movements were observed when the subjects moved as if directing traffic. The hand was moving in smooth circles - in case (1) the circles were big, in case(2) they were smaller. The velocity remains low and constant during mass-spring movements, going to 0 only at the end of the movement. The other two plots show velocity profiles of movements during reaching and striking. The ballistic movements have a characteristic "bell" shaped profile. The secondary bells occurring in the case of reaching correspond to the retraction phase of the movement. As there is higher acceleration and deceleration during striking compared to reaching, the bells in the profiles in the case of striking are more convex than those for reaching.

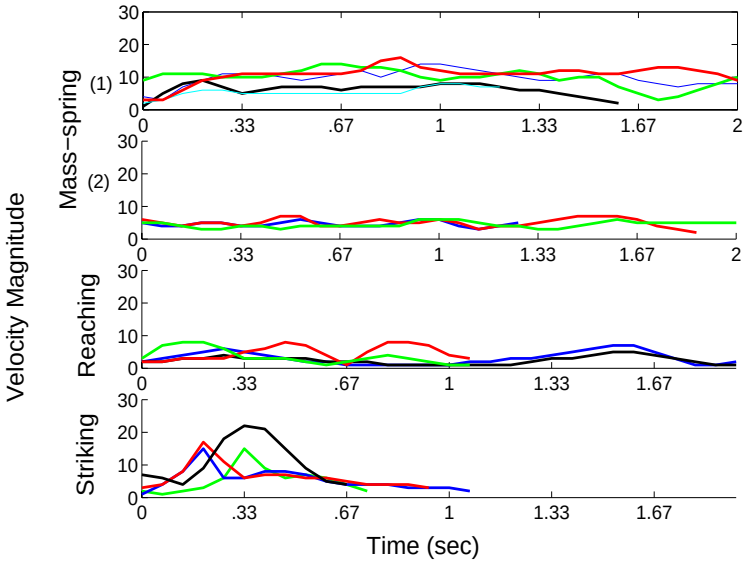


Figure 2.3: Examples of velocity profiles for mass-spring and ballistic movements.

The nature of a ballistic movement is determined by the dynamics. For ex-

ample, reach movements have low acceleration and deceleration, strike and throw movements have high acceleration and deceleration. There is also the possibility of yanking - this has high acceleration, the deceleration may vary. Figure 2.4 illustrates this with a schematic.

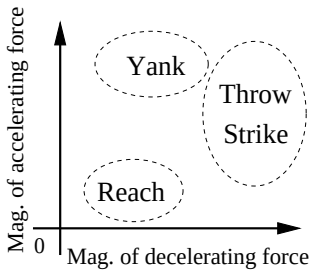


Figure 2.4: Varying the parameters of the ballistic movement model produces different types of movements: low acceleration and deceleration for reach, high acceleration and deceleration for throws and strikes, and high acceleration for yanking.

2.3 Empirical Studies of Reach Movements

Studies of reaching movements have shown that the shape of the "bell" varies considerably depending upon the task requirements [51]. For instance, in reach-to-grasp movements, when the object is small or fragile, the deceleration phase has a longer duration. One possible explanation is that this gives more time for precise homing on to the target. There have been subsequent studies with more detailed analysis, e.g., reaches with rotating torso movement [60], reaches with single step and free torso movement [27]. These studies indicate that humans plan reach movements in an extrinsic (world-centric) reference frame, rather than in a frame fixed to their torso or head. Some important observations are:

- Given the starting and target position of the hand for a reach movement, and

in the absence of constraints such as via points and obstructions, the hand typically follows a near straight-line path with a "bell" shaped velocity profile in extrinsic coordinates. The "bell" feature is present even in the case of substantial torso rotation and single-step leg motion.

However, studies of reach movements in which the targets were placed at the extremes of the arm's work-space, e.g. [35, 85], etc., indicate that the paths can also have substantial curvature. These curvatures are reported to be consistent for repeated movements to the same target.

- The hand's path in extrinsic coordinates is relatively unaffected by variations in the movement's pace.

- Viewed in the joint angle space, the coordination of various joint movements varies with changes in the pace of the movement. That is, the timing of shoulder and elbow flexion/extension changes when the overall speed is varied.

- In a trunk-based reference frame, the velocity profiles of the hands are multi-peaked and exhibit greater variability. This is in contrast to the "bell" shape of the profiles in extrinsic coordinates.

- When a person is adept at executing a movement, the whole body moves in synchrony. The various body-parts such as the head, torso and hands start and stop motion in a coordinated manner. This has been observed even in case of periodic movements such as walking, e.g., [41].

There are also studies of human perception of movement, e.g., by Johansson [41]. Here, human observers recognize various human gaits even though only the 2D positions of various joints and extremities of subjects were made observable.

2.4 Dynamical Models of Reach Movements

Here, trajectory refers to both the spatial path and the velocity profile of the hands during movements. Several models have been proposed for human control of hand movement for reaching - we concentrate on the computational models. These can be open-loop - the trajectory is generated given the initial and final position, velocity and acceleration, e.g. [28, 85, 80]. or closed-loop - the control is continuously adjusted according to errors in limb propulsion and target perturbation, e.g., [37].

2.4.1 Minimum Jerk Model (MJM)

Flash and Hogan [28] proposed that, as practised movements are smooth, minimizing the mean-square jerk could be one of the criteria used by humans for planning trajectories. Using the calculus of variations, they show that for this minimization, the trajectory in each coordinate should follow a 5^{th} order polynomial. By setting the initial and final velocity and acceleration to 0, the model was able to replicate the near straight-line paths and "bell"-shaped velocity profiles observed for short reach movements in humans. The model is limited as it does not take into account dynamical factors like gravity, arm lengths, etc. Moreover, several studies have cited the pronounced curvature of paths followed by hands when reaching in

certain directions, e.g., [35, 85]. In spite of these inherent limitations, the model has been shown to predict intermediate trajectories with reasonable accuracy [28].

2.4.2 Minimum Torque Change Model (MTCM)

This was proposed by Uno et al. to address some of the limitations of MJM [85]. Here, the objective is to minimize the change in the torques acting on different joints of the arm. This model takes into account dynamical factors like mass, moment of inertia of limbs, gravity, joint viscosity, etc., and the fact that the human arm is multi-jointed. It is shown to replicate several features of reach movement paths such as curvatures of the paths for movements in certain directions. The model presented in [85] is for 2D planar movements and does not include the torso. Moreover, it involves non-linear optimization and knowledge of the subject's initial and final joint (elbow) configuration. The advantage is that in addition to the hand, the model also predicts the elbow's trajectory.

2.4.3 Minimum Peak Energy Model (MPEM)

Donder's Law states that for every gaze direction, there is a unique orientation of the eyes w.r.t. the head [80]. Applied to arm movement, the analogy would be that for every target position of the hand in the 3D space around the subject, there exists a unique configuration of the arm at the end of the movement. However, experimental observations indicate that the terminal arm configuration is not a unique function of the target hand position but also depends upon the starting

position [80]. This is in addition to intuitively obvious factors such as the required final wrist orientation, physical constraints like limb-lengths, etc. Grounded on these experiments, a model for hand movement based on the minimization of peak energy expended during motion is proposed in [80]. The arm is allowed to move in 3D, but the torso is assumed to be immobile. The model predicts the arm's final configuration given the initial arm configuration and the final position of the hand. The results match well with experimental observations. However, it is not clear if the hand's predicted *path* would be similar to that of actual movements. The reason is that MPEM assumes that all joints will reach peak velocities at the same time; whereas it has been observed that joint movement coordination varies with change of pace and load characteristics.

2.4.4 Minimum Jerk Model with Feedback

Hoff and Arbib extended MJM with feedback control to accommodate errors in the hand's propulsion, noisy observations, target perturbation in mid-flight, etc. [37]. The predicted results match well with experimental observations. The authors state that for well-practised movements, and in the absence of target perturbation, the propulsion speeds are usually so high that a large part of the spatial path is covered with little or no feedback; the control is more akin to open-loop predictive systems.

2.5 Summary of Psycho-kinesiological Observations

Empirical studies suggest that detecting and tracking just the hands and feet of subjects might be adequate for recognizing human movements such as reach, strike, walking etc. Moreover, the velocities of the hands have high correlation with velocities of other body-parts such the arm, head, torso, etc. This is advantageous from an image processing perspective as we can avoid the complex task of estimating the complete configurations of a subject's joints. Low-level motion features computed from the whole figure of the subject may be used to implicitly represent the hand's velocity. This will be described in detail in the chapter on video-based analysis. In addition, for ballistic movements, when the hands are observed in extrinsic coordinates, the spatial paths and velocity profiles have simple structure. This is also advantageous because analysis in a world-centric reference frame is simpler as it can be fixed to the static camera. In contrast, a torso-centric reference frame would require accurate tracking of the torso orientation, which is a complex task.

Although the Minimum Jerk Model (MJM) ignores several important dynamical properties, it still has some advantages over the other models for visual movement analysis:

- It does not require estimation of the subject's joints - a hard problem for computer vision systems [31, 53].

- Although the other models include elbow and shoulder information, they still ignore torso and whole body (stepping) movements. These are important components of movements observed in everyday life. Hence, it is not clear

how much of a practical advantage the other models would have over MJM.

- MJM uses a 5^{th} order polynomial for predicting the trajectory. In contrast, MTCM involves non-linear optimization. MPEM, claimed to be computationally simpler than MTCM, does not predict the hand's actual trajectory. The Hoff-Arbib model might be unnecessarily complex for the practised movements considered in the present study.

Chapter 3

A Bayesian Model for Recognizing Ballistic Movements

Visual recognition is complex due to the presence of noise and ambiguity in the extracted features such as edges, depth, texture, pose, motion in terms of optical flow, etc. Bayesian probabilistic inference provides a principled framework for handling uncertainty in recognition. Consequently, it has been extensively employed in computer vision. This chapter presents a Bayesian framework for recognizing ballistic movements that incorporates the psychological observations described in Chapter 2. The approach is illustrated with experiments on human motion capture data - which has the advantage of low noise. Chapter 4 introduces a video-based recognition system employing this framework.

3.1 A Bayesian Model for Ballistic Movements

Human activity may be modelled as a sequence of movements executed to interact with objects and the environment. To make recognition tractable, vision approaches assume the conditional independence of movements. I.e. each movement is considered to be independent of past and future movements given the context provided by the activity, and the states of the subject at the start and end of the movement. Ballistic movements such as reaches and strikes are atomic by nature. Once started, they run their course to the end of the movement. Thus, the inde-

pendence assumption is well suited for recognizing them. The equivalent Bayes net is shown in Layer I of the model shown in Figure 3.1. Layer II of the model consists of the dynamics, B_i, that control the trajectory of the hand during a movement. Layer III consists of observations of image features for pose and motion estimation. The present study focuses on recognizing individual movements, i.e. Layers II and III.

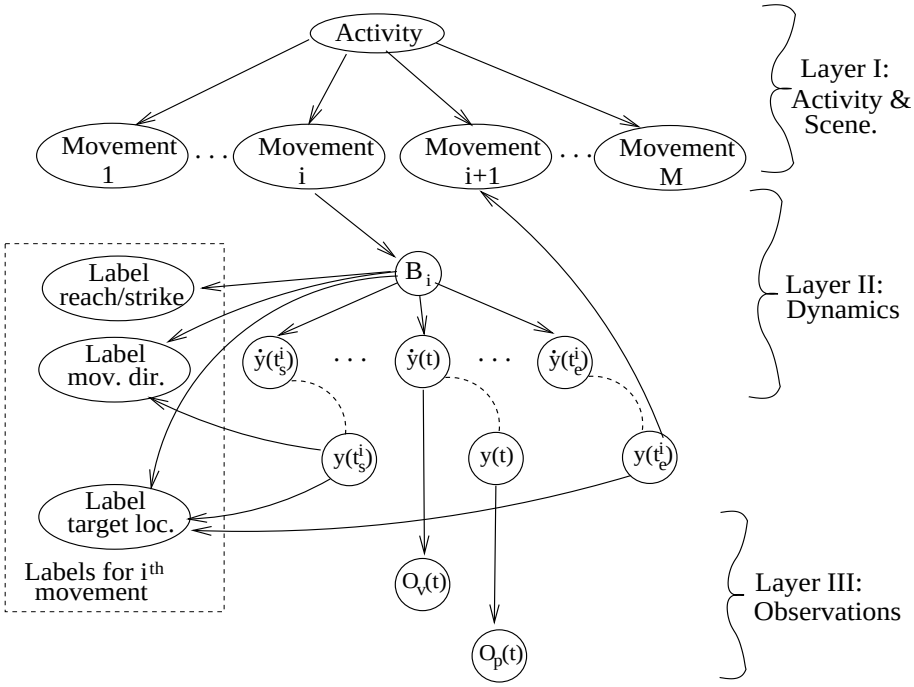


Figure 3.1: Bayes net for modelling ballistic movements. This is similar to the structure proposed by Bregler [16].

3.1.1 Model for the Hand's Trajectory

The Minimum Jerk Model (MJM) minimizes the rate of change of force applied to the hand - the intuition being that efficient movements are smooth [28]. Let $\mathbf{z}(t) = [z_1(t)\ z_2(t)\ z_3(t)]^{\mathrm{T}}$ be the hand's coordinates in 3D world coordinates. Then,

the trajectory minimizes

$$J = \frac{1}{3}\int_{t_s}^{t_e} \left(\left(\frac{d^3 z_1}{dt^3} \right)^2 + \left(\frac{d^3 z_2}{dt^3} \right)^2 + \left(\frac{d^3 z_3}{dt^3} \right)^2 \right) dt \tag{3.1}$$

It can be shown using Calculus of Variations that minimizing the functional $J(.)$ is equivalent to constraining z_1, z_2 and z_3 to be 5^{th} order polynomials in time t. Even though the hand's trajectory is a high order polynomial, the path followed by the hand during ballistic movements is relatively simple - closely corresponding to straight lines. The higher order terms in the function are "taken up" in the high acceleration and deceleration involved in the movements. Let $\vec{\tau}(t)$ denote a column vector such that $\vec{\tau}(t) = [1\ t \ldots t^5]^{\mathrm{T}}$. Let the duration of the i^{th} movement be $[t_s^i, t_e^i]$. Its trajectory is given by

$$\mathbf{z}(t) = \hat{A}_i \vec{\tau}(t - t_s^i) \quad \text{where} \quad t \in [t_s^i, t_e^i] \tag{3.2}$$

Differentiating both sides w.r.t. time gives the velocity

$$\dot{\mathbf{z}}(t) = \hat{B}_i \vec{\tau}(t - t_s^i) \quad \text{where} \quad t \in [t_s^i, t_e^i] \tag{3.3}$$

Here $\hat{B}_i \vec{\tau}(t) = \hat{A}_i \frac{d}{dt}\vec{\tau}(t)$. The dynamics of the i^{th} movement in 3D world coordinates is represented by $\hat{B}_i$.

Projective Transformation: Let $\tilde{\mathbf{z}}(t)$ denote the hand's position in homogenous coordinates - $\tilde{\mathbf{z}}(t) = [z_1(t)\ z_2(t)\ z_3(t)\ 1]^{\mathrm{T}}$. Let $\tilde{A}_i$ correspond to the homogenous version of $\hat{A}_i$, i.e.

$$\tilde{A}_i = \begin{bmatrix} & \hat{A}_i & \\ & & \\ 1 & \mathbf{0}^{\mathrm{T}} & \end{bmatrix}.$$

Therefore, for the i^{th} movement, we have

$$\tilde{\mathbf{z}}(t) = \tilde{A}_i \vec{\tau}(t - t_s^i) \tag{3.4}$$

Let P denote the projection matrix for the camera. Let the projected trajectory in homogenous coordinates be $\tilde{\mathbf{y}}(t) = [\tilde{y}_1(t)\ \tilde{y}_2(t)\ w(t)]^{\mathrm{T}}$. It is given by

$$\tilde{\mathbf{y}}(t) = P\tilde{\mathbf{z}}(t) = P\tilde{A}_i \vec{\tau}(t - t_s^i)$$

Thus, the hand's trajectory remains a 5^{th} order polynomial under projection in homogenous coordinates.

Let $\mathbf{y}(t)$ denote the hand's position in image coordinates. If the change in the hand's depth w.r.t. the camera is small compared to its distance from the camera, then $w(t)$ can be assumed to be constant for the duration of a movement. This results in

$$\mathbf{y}(t) \approx \lambda_{\text{depth}}[\tilde{y}_1(t)\ \tilde{y}_2(t)]^{\mathrm{T}} \tag{3.5}$$

Thus, the projection of the trajectory on the image plane can be closely approximated by a 5^{th} order polynomial in time. Under similar assumptions, it can be shown that the projections of the hand's velocities on the image plane are 4^{th} order polynomials in time. Let $\dot{\mathbf{y}}(t)$ denote the projected velocity

$$\dot{\mathbf{y}}(t) = B_i \vec{\tau}(t - t_s^i) \tag{3.6}$$

B_i determines the dynamics of the i^{th} movement on the image plane. Due to the assumption of ballistic dynamics, B_i is constant for the duration of the i^{th} movement. Therefore, the velocities, $\dot{\mathbf{y}}(t)$, are mutually independent given B_i. Layer II in

Figure 3.1 shows the equivalent Bayes net structure. B_i models only the hand's velocities which may be estimated using low-level motion features such as optical flow. This enables recognition without explicit tracking of the poses. As $\mathbf{y}(t)$ and $\dot{\mathbf{y}}(t)$ are the position and velocities at time t, they are implicitly linked by time (shown with dashed lines).

3.1.2 Observation of the Hand's Position and Velocity

Accurately tracking the subject's body during movements is perhaps the most challenging aspect of action recognition [53]. Capitalizing on the ballistic nature of reach and strike movements enables recognition without explicitly tracking the poses.

Hand's position vs. its velocity: Psychological studies provide two useful observations:

1. Studies of reaches involving torso movement and stepping, e.g., [60, 27], indicate that the whole body moves in synchrony with the hands during ballistic movements. The start and stop of the hand's motion, and its velocity are reflected in the velocities of other body-parts.
2. The hand usually follows a simple path during ballistic movements, closely resembling straight lines and low curvature 3D circular arcs [79, 28, 85].

As a result, low-level motion features such as optical flow and silhouette deformation computed over the whole figure of the subject have high correlations with the hand's velocity. Figure 3.2 illustrates this for optical flow. Thus, the hand's ve-

locity is approximated without tracking the pose. Moreover, the high movement speeds, while making position estimation difficult, give distinctive signatures to the motion features. Thus, position information is reliable during the start and end of movement, and velocity observations are more robust during mid-flight.

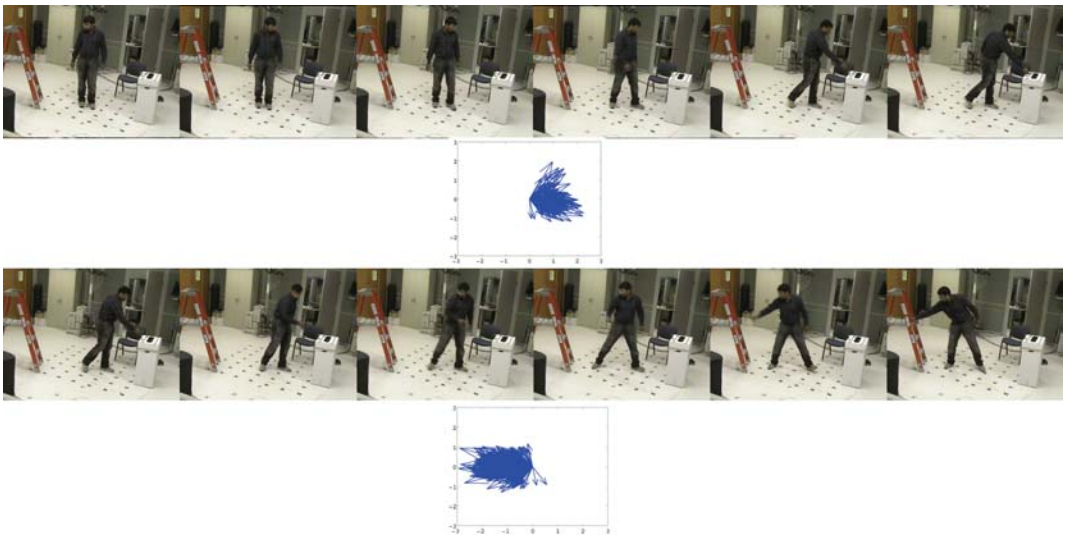


Figure 3.2: Every third of a sequence of 41 frames is shown depicting two movements. The optical flow vectors computed computed on the person's figure during each movement segment are below it (all vectors have been translated to the origin). A majority of the flow vectors point in the direction of movement. (Best viewed in color.)

Observation Model: For video-based analysis, the position observations, O_{p}, represent the subject's pose. The velocity observations, O_{v}, consist of optical flow, silhouette deformation and frame differences. This is described in Chapter 4. We make the standard conditional independence assumptions for the observations given the position and velocity, i.e.

$$\begin{aligned} p(O_{\mathrm{p}}|\mathbf{y}) &= \prod_{t=t_s^i}^{t_e^i} p(O_{\mathrm{p}}(t)|\mathbf{y}(t)) \\ p(O_{\mathrm{v}}|\dot{\mathbf{y}}) &= \prod_{t=t_s^i}^{t_e^i} p(O_{\mathrm{v}}(t)|\dot{\mathbf{y}}(t)) \end{aligned} \tag{3.7}$$

The Bayes net structure is shown in Layer III of Figure 3.1.

3.1.3 Overview of Recognition: Label Inference

Continuous sequences are segmented into individual ballistic movements by employing the property that the dynamics in terms of B_i remains constant for the duration of a ballistic segment. This is performed using weighted least squares estimation and dynamic programming (c.f. Section 3.3). Reach/Strike labels are inferred by modelling the statistics of the dynamics (c.f. Section 3.4). Qualitative labels of the movement's direction and target location are computed using the starting pose as the reference frame (c.f. Section 3.5). The nodes corresponding to the labels are shown in the dashed rectangle in Figure 3.1.

3.2 Related Work

There have been a large number of studies on action recognition - see [53, 31] for comprehensive surveys.

Bregler presented an approach for recognizing complex actions as a sequence of simpler actions [16]. At the lowest level, actions are considered to be atomic, called "movemes". It is interesting to note that actions having ballistic movement are atomic by nature. Our work can be considered as an approach for representing and recognizing movemes that are ballistic. Closely related, there are studies using Switching Linear Dynamical Systems (SLDSs) for characterizing human movement e.g., [68]. In addition, many approaches use the dependencies between the move-

ments of different body-parts, e.g., [7].

Wilson and Bobick proposed Parametric Hidden Markov Models (P-HMMs) to handle variability in gestures [91]. P-HMMs would need a sufficient variety of training examples to generalize over all possible target locations. However, as they model the trajectory of movement, their approach can be used for recognizing different mass-spring movements like communicative gestures. In this respect, our work and P-HMMs complement each other.

Rao et al. proposed a scheme for segmenting human movement sequences based on the spatio-temporal curvatures of the hands' trajectories [67]. Weinland et al. segment continuous movement sequences using Motion History Volumes computed using 3D reconstruction[89]. The temporal segmentation in our approach uses single camera-view video and does not require tracking the hands.

State-of-the-art sub-space methods, e.g., [94, 77], have been developed to perform recognition robust to camera viewpoint and stylistic variation. Even for a stationary camera, two reach movements can have very different body-part trajectories if their target locations differ. Therefore, recognizing them involves generalizing over the dynamics in addition to the viewpoint. Our approach contributes in this direction. A possible area of future study would to be to employ approaches such as [77] to explore the variation of matrix B_i w.r.t. subtle movement styles.

It is possible to extend the approach by including object interaction [59] - this would help differentiate between actions such as "picking up" and "putting down". The proposed approach analyzes each ballistic movement independent of past and future movements. It is possible to link the dynamics of ballistic movements with

HMMs, generative grammars, etc. See [53] for a survey. In addition, spatial context and geometry of the scene have been used to aid object recognition, e.g., [83]. This study focuses only on the recognition of individual ballistic movements - linking it with temporal, object and scene-geometry context is an area of future research.

3.3 Segmenting Movements

A continuous movement sequence is segmented such that the dynamics, B_i, within each subsequence is constant. The B_i's are estimated using weighted least squares, and Dynamic Programming is used to efficiently compute the optimal segmentation. Let the sequence be of time duration $[0, T]$. Let χ denote a partitioning of the sequence into n segments, $\chi = \langle \chi_0 = 0, \chi_1, \ldots, \chi_n = T \rangle$. The start of the i^{th} movement is $t_s^i = \chi_{i-1}$, and end is $t_e^i = \chi_i$. The likelihood of the segmentation given the velocity observations, $p(\chi|O_{\text{v}})$ is modelled as $p(\chi|O_{\text{v}}) = p(B_1^* \ldots B_n^*|O_{\text{v}})$, where B_i^* is the optimal dynamics for the i^{th} partition given the observations. By the conditional independence assumption

$$\begin{aligned} p(B_1 \ldots B_n|O_{\text{v}}) &= \prod_i^n p\left(B_i \left| O_{\text{v}}(t_s^i) \ldots O_{\text{v}}(t_e^i)\right.\right) \\ &= k \prod_i^n p(O_{\text{v}}(t_s^i) \ldots O_{\text{v}}(t_e^i)|B_i)\ p(B_i) \end{aligned} \tag{3.8}$$

Here, k is a constant independent of the partitioning, and $p(B_i)$ is the prior on the dynamics. The prior enforces constraints such as starting and ending velocity magnitudes should be close to 0. $p(O_{\text{v}}(t_s^i) \ldots O_{\text{v}}(t_e^i)|B_i)$ is the conditional probability of the velocity observations given the dynamics. Given its segment boundaries $[t_s^i, t_e^i]$, the goodness of the i^{th} segment is independent of the rest of the segmentation.

Due to this Markovian property, the optimal partitioning, χ^*, can be efficiently computed using Dynamic Programming. Figures 2.1, 3.2, 4.5 and 4.6 show examples of obtained segmentations. Details of $p(O_{\mathrm{v}}(t_s^i) \ldots O_{\mathrm{v}}(t_e^i)|B_i)$, the DP algorithm and quantitative results are described in Chapter 4.

3.4 Classifying Movements based on Dynamics

The nature of a ballistic movement is determined by the dynamics. For example, reach movements have low acceleration and deceleration, strike and throw movements have high acceleration and deceleration. There is also the possibility of yanking - this has high acceleration, the deceleration may vary. Figure 2.4 illustrates this with a schematic. The reach vs. strike labels are computed by modelling the statistics of O_{v}. We use a boosting framework to get the MAP label estimate [1]. Chapter 4 presents details of the features employed for video-based recognition and the experimental results.

3.5 Computing Labels for Movement's Direction and Target Location

After classifying a ballistic segment into reach or strike, the target's location and the direction of movement are described using qualitative labels. To be mutually consistent, the labels of different movements must be computed in appropriate reference frames. The reason is that the reference frame is the principal factor determining invariants during recognition. For example, when recognizing arm gestures,

the movements must be recognized with respect to the person's body. In contrast, pointing gestures (indicating the direction to proceed) should be recognized in a world centric reference frame. In general, there are at least three possibilities for the reference frame [79, 60]:

1. World-centric.
2. Body-centric: e.g., the frame could be fixed to the torso, the hand, or the head's gaze-direction.
3. Fixed to the object being manipulated in a movement.

In our approach, the reference frame for a movement is fixed to the person's pose at the start of the movement. This provides two advantages:

- As the frame is fixed to the person's pose, the movement's labels are computed with respect to the person's perspective at the start of the movement. Thus the labels are mutually consistent regardless of the person's position relative to the camera.
- Because the frame is constant for the duration of a movement, it is inertial. Psychological studies indicate that the velocities of the body-parts have greater consistency when viewed in an external fixed reference frame [60]. This provides robustness during recognition.

The label, l, for each movement is a 3-tuple $\langle l_a, l_e, l_d \rangle$:

1. l_a is the azimuthal location of the target. $l_a \in L_a$ ={front, back, left, right and center}.

2. l_e is the elevation location of the target. $l_e \in L_e$ ={ankle-level, knee-level, waist-level, chest-level and above-shoulder}.
3. l_d is the direction of movement. $l_d \in L_d$ ={forward, backward, leftward, rightward, upward and downward}.

See Figure 3.3 for an illustration.

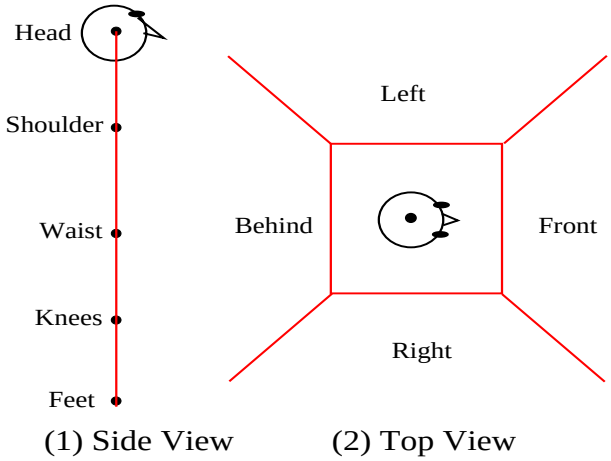


Figure 3.3: Spatial quantization of the space around the person for computing movement labels.

Target Location: For reach movements, the target is located at the end of the reach. For strikes, the target is located at the position of highest velocity of the hand. Let $\mathbf{y}_{\text{target}}$ denote the location of the target in the image. If a ballistic segment with time-interval $[t_s, t_e]$, has been classified as a reach then $\mathbf{y}_{\text{target}} = \mathbf{y}(t_e)$. If the ballistic segment has been classified as a strike then

$$\mathbf{y}_{\text{target}} = \mathbf{y}(t_{\max}) \quad \text{where} \quad t_{\max} = \arg\max_{t=t_s}^{t_e} h(O_{\text{v}}(t))$$

where $h(.)$ is the reach/strike classifier's confidence function.

Spatial Context for Labelling: The subject's pose at the start of the movement is represented by the subject's silhouette and the head's gaze-direction

at the start of the movement - O_{p}, c.f. Chapter 4. It is used to provide context to the target's location and the direction of movement. The label for the movement is computed based on this context. Consider a ballistic segment with time-interval $[t_s, t_e]$. Let the subject's starting pose be $O_{\mathrm{p}}(t_s)$, and the hand's target position in the image be $\mathbf{y}_{\mathrm{target}}$. The label for the target's position depends upon the location of $\mathbf{y}_{\mathrm{target}}$ relative to $O_{\mathrm{p}}(t_s)$.

Bayesian inference is employed to compute the label for each movement. Let $p_a(l_a|\mathbf{y}_{\mathrm{target}}, O_{\mathrm{p}}(t_s))$ denote the likelihood of label l_a for the azimuthal position, given $\mathbf{y}_{\mathrm{target}}$ and $O_{\mathrm{p}}(t_s)$. Noise present in the video causes ambiguity in the estimation of the pose and the hand's target position. Therefore, the probability of l_a is computed by marginalizing over them:

$$p(l_a) = \sum_{O_{\mathrm{p}}(t_s)} \sum_{\mathbf{y}} p_a(l_a|\mathbf{y}, O_{\mathrm{p}}(t_s)) p(\mathbf{y} \text{ is target } |O_{\mathrm{p}}(t_s)) p(O_{\mathrm{p}}(t_s)) \tag{3.9}$$

$P(O_{\mathrm{p}}(t_s))$ denotes the probability of the pose observations. $P(\mathbf{y}$ is target$|O_{\mathrm{p}}(t_s))$ is the probability of the target of the movement to be located at point $\mathbf{y}$ in the image, given the starting pose.

The probabilities for the elevation labels are formulated similar to Eq.(3.9).

$$p(l_e) = \sum_{O_{\mathrm{p}}(t_s)} \sum_{\mathbf{y}} p_e(l_e|\mathbf{y}, O_{\mathrm{p}}(t_s)) p(\mathbf{y} \text{ is target } |O_{\mathrm{p}}(t_s)) p(O_{\mathrm{p}}(t_s)) \tag{3.10}$$

For computing direction labels, the target's location is replaced by B_i.

$$p(l_d) = \sum_{O_{\mathrm{p}}(t_s)} \sum_{\mathbf{y}} p_d(l_d|B_i, O_{\mathrm{p}}(t_s)) p(B_i|O_{\mathrm{p}}(t_s)) p(O_{\mathrm{p}}(t_s)) \tag{3.11}$$

The final label for each ballistic segment is computed as the maximum a posteriori probability estimate. Figures 2.1 4.6 and 4.5 show some examples of computed

labels.

Chapter 4 describes the image and motion features used for computing the pose and velocity observations, the training, and the inference algorithm employed for recognition. Quantitative results on video analysis are also presented.

The next section illustrates the Bayesian framework with experiments on motion capture data. Here, the position and velocity of the hands and other body-parts are directly observable with relatively low noise. Therefore, these experiments are intended as a "sanity check" of the framework.

3.6 Analysis of Motion Capture Data

We analyze marker-based motion capture data from the CMU MoCap database [4]. Results of this work were published in [63]. In motion capture, special markers are attached to different parts of the subject's body such as the head, hand, elbow, etc., and tracked with high performance cameras. The pose estimation is very accurate with relatively low noise in the markers' localization. Although ideal for illustrative purposes, this methodology is intrusive and not practical for applications such as surveillance. Each motion capture sequence consists of a sequence of 3D locations of various markers, available at 60Hz to 120Hz. To emulate typical video recording frame rates, the sequences were down-sampled to 15Hz before analysis in the experiments. Figure 3.4 shows skeleton-plots of the markers for two segments of strike movements, and the corresponding labels computed by the approach.

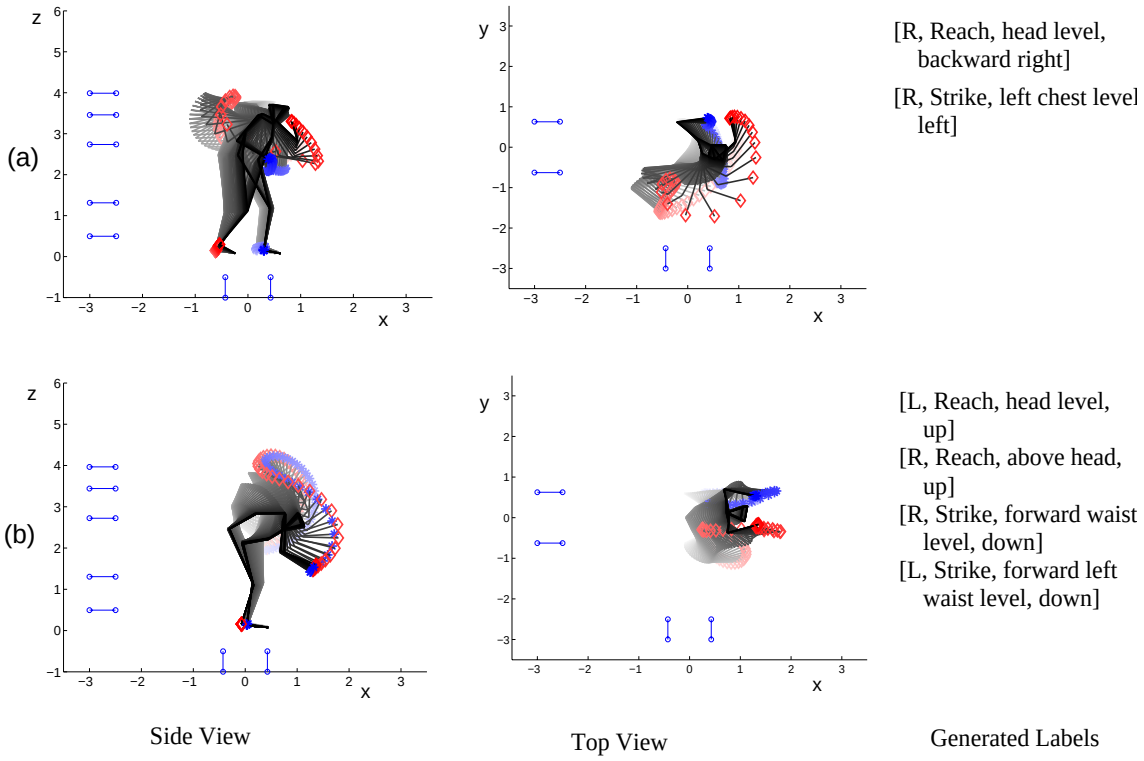


Figure 3.4: Two instances of striking: (a) slapping someone's back, (b) banging on a table with both hands. In both cases, the subjects first draw back their hands before striking. Skeletons at different time instants are plotted - older ones have faded colors. Red diamonds correspond to the right hand and leg; blue asterisks are for the left hand and leg. The blue stubs placed along the axes mark front/back, left/right, and height reference points for the subjects. The labels generated by the proposed system are listed alongside in the order generated.

3.6.1 Model for Ballistic Force Actuation

Consider the following simple model for force actuation during a ballistic movement. Let m be the mass of the body part, f^+ the accelerating force and f^- be the decelerating force. Starting at time $t = 0$, f^+ acts on m for time t_1. After this, the body part moves ballistically for time t_2. Finally, the deceleration force, f^-, acts on m for time t_3. As the body part comes to a near stop at the end of a ballistic movement like reach, etc., f^+ and f^- oppose each other. For simplicity, we ignore

gravitational force. Let $T = t_1 + t_2 + t_3$ be the total duration of the movement and D be the total distance. Figure 3.5 shows a schematic of the velocity profile. The plan for the movement, called the execution plan, would be specified by t_1, t_2, t_3, f^+ and f^-. Depending upon the values of f^+ and f^-, a ballistic movement could act as a reach, strike, etc.

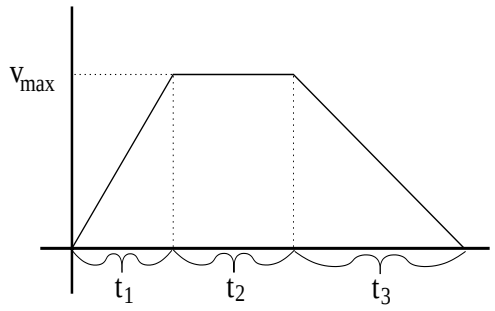


Figure 3.5: Schematic of the velocity profile during a ballistic movement.

For each type of movement, the motion parameters are further tuned to suit the task at hand. For example, during reaching, if the target is small or fragile, t_3 is considerably longer and f^- is relatively low. This increases the precision in homing onto the target and provides more time for adjusting the wrist and finger positions during the final approach [51].

The movement parameters are not observable from the hand/foot trajectories. Let $v(t)$ be the velocity magnitude of the hand/foot during a movement. The movement's dynamics can be described implicitly in terms of the following observable quantities:

1. The peak velocity reached during the movement - $v_{\max}$

2. The second derivative of the velocity at the location of the peak - $\ddot{v}(t_p)$.

3. The total time duration of the movement, T.

4. The total distance travelled during the movement, D.

3.6.2 Segmentation of sequences into ballistic movements

A continuous motion capture sequence is segmented into individual ballistic movements based on the dynamics of the hands. Ideally, the velocity profile of each segment would have a monotonically non-decreasing phase followed by a monotonically non-increasing phase. However, noise in the observations may cause false extrema in the velocity profile. Instead of explicitly modelling the noise, we treat this as a problem of classifying local minima that actually demarcate ballistic subsequences from those caused by noisy observations. Each local minima was characterized by the decelerating impulse preceding it, the time duration of this impulse, the speed at the minima, the accelerating impulse following it and its duration.

In addition to segments exhibiting motion, there are segments with little or no motion. These are characterized by their maximum velocities being below a certain threshold. Given confidence values for each time instant to be a starting, ending or negligible movement, we compute the most likely segmentation of the capture sequence using Dynamic Programming.

Let $p^*(t)$ denote the likelihood of segmentation such that the last segment ends at t. Let $\alpha_t(t_s)$ be the likelihood for the most likely segmentation whose last segment starts at t_s and ends at t. Let $\beta_t(t_s)$ be the likelihood for the most likely segmentation whose last segment starts at t_s and *continues beyond* t. Let $s(t)$ be the likelihood

for t to be a start of a ballistic movement, and $e(t)$, for t to be an ending. Let $\delta_t(t_s)$ be the likelihood for the most likely segmentation such that the last segment has negligible movement, starts at t_s and ends at t. A negligible segment must be preceded by a non-negligible segment. We have the following recursive relations:

$$\begin{aligned}
\beta_t(t_s) &= \begin{cases} p^*(t-1)\ s(t) & t_s = t \\ \beta_{t-1}(t_s)(1-e(t)) & t_s < t \end{cases}; \\
\alpha_t(t_s) &= \begin{cases} p^*(t-1)\ s(t)e(t) & t_s = t \\ \beta_{t-1}(t_s)e(t) & t_s < t \end{cases}; \\
v_t^*(t_s) &= \begin{cases} v(t) & t_s = t \\ \max\left(v_{t-1}^*(t_s), v(t)\right) & t_s < t \end{cases}; \\
u(t) &= \max_{t'=0}^{t-1} \alpha_t(t'); \\
\delta_t(t_s) &= u(t_s)\Psi\left(v_t^*(t_s)\right); \\
p^*(t) &= \max_{t'=0}^{t-1} \left(\max(\alpha_t(t'), \delta_t(t'))\right)
\end{aligned} \tag{3.12}$$

Here $\Psi(v) = [v \leq 2]$ - it maps velocity magnitudes to likelihoods of being negligible. The recursive functions can be computed with linear time and space complexity[1]. For the first step in the computation, i.e. for $t = 1$, we keep $p^*(0) = 1$ and $v^*(0) = 0$. For an optimal segmentation whose last segment starts at t_s, let $prev_t(t_s)$ point to the segment preceding the last segment.

$$prev_t(t_s) = \begin{cases} \arg\max_{t'=0}^{t-1} \left(\max(\alpha_t(t'), \delta_t(t'))\right) & t_s = t \\ prev_{t-1}(t_s) & t_s < t \end{cases} \tag{3.13}$$

[1]The time complexity is made linear by assuming that valid segments cannot be greater than a certain length (2 secs.).

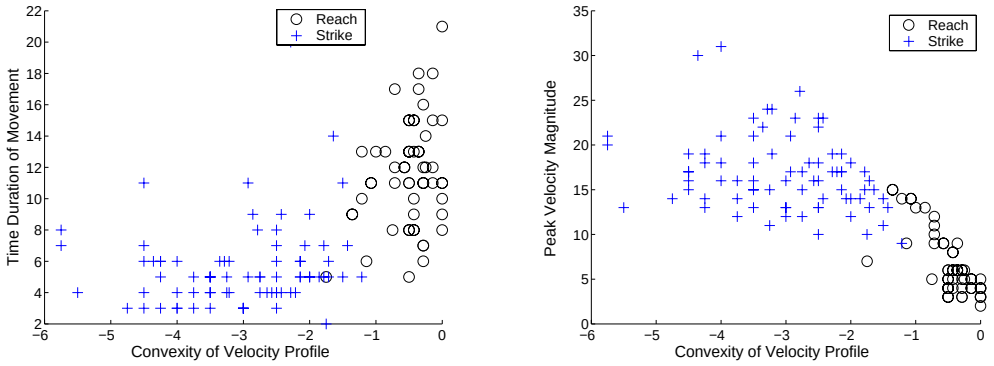
Let $\phi_s(i)$ and $\phi_e(i)$ denote the start and end of the i^{th} segment in the optimal segmentation. After the set of relations (3.12) and (3.13) are computed for $t = 1 \ldots T$, the optimal segments are recovered recursively as:

$$\begin{aligned}
\phi_s(n) &= \begin{cases} \arg\max_{t=0}^{T-1} \left[\alpha_T(t), \delta_T(t)\right] & n = N \\ prev_T(\phi_s(n+1)) & n < N \end{cases} ; \\
\phi_e(n) &= \begin{cases} T & n = N \\ \phi_s(n+1) & n < N \end{cases}
\end{aligned} \tag{3.14}$$

Here N is the number of segments in the optimal segmentation. (This need not be known a priori and is simply used to describe the computation.) The obtained segments are post-processed to eliminate irrelevant movements. Only movements in which the hand moves by a distance greater than the length of the subject's forearm are considered relevant. In addition, the spatial quantization described previously is used to define a volume around the waist of the subject in which the hands are usually located when at rest. Movements with target locations in this volume are considered to be irrelevant.

3.6.3 Classification into reaches and strikes

To illustrate the efficacy of the ballistic dynamics parameters, Figure 3.6 shows scatter-plots of the $\ddot{v}(t_p)$ vs. T, and $\ddot{v}(t_p)$ vs. $v_{\max}$, for the reach and strike segments. As strike movements have greater acceleration and deceleration, their velocity peaks are more convex (more -ive). Moreover, they are faster, so their time durations are small and the maximum velocities are higher than those of reach movements. There is a significant separation in the distributions of the two types of movements.



(a) (b)

Figure 3.6: Scatter-plot of (a) $\ddot{v}(t_p)$ vs. T, (b) $\ddot{v}(t_p)$ vs. $v_{\max}$.

An SVM was used to distinguish between reaching and striking [2]. For the experiments 64 samples collected for reaching and 83 for striking were used. Each sample was represented by a 3D vector consisting of $\ddot{v}(t_p)$, T and $v_{\max}$. The experiments consisted of 100 trials, in each trial a portion of the data was randomly chosen for training and the rest was used for testing. Table 3.1 shows the classification results in terms of the mean and variance of the classification accuracies. The accuracies are high and their variance is low, indicating that the features adequately characterize the ballistic nature of reaching and striking movements, and that the distributions are stable.

	Mean	Std. Dev.
Reach	0.9478	0.0449
Strike	0.9690	0.0377

Table 3.1: Means and standard deviations of the classification accuracies for reaching vs. striking over 100 trials of SVM training and testing.

3.6.4 Labels for Movement's Target location and Direction

Computing labels for the hand's target location and the movement's direction is relatively simple due to highly accurate body-part localization in the motion capture data. The objective is verify whether the labels computed using the approach are coherent with visual perception of the movements.

3.6.4.1 Reference Frame for Describing Movement

We define the movement's coordinate system as the subject's reference frame at the time the movement commences. As this is the time and location when the subject planned and began execution, the generated description would be consistent not only with his/her viewpoint, but also with similar movements executed at other times and locations. A 3D orthogonal coordinate system is used - the x-axis is along the front-back direction, the y-axis is along the left-right direction, and the z-axis is always vertical. The origin is kept on the ground plane. The azimuthal orientation and the x and y coordinates of the origin are computed using 4 motion-capture markers fixed to the subject's waist. See Figure 3.7(a) for an illustration. Let $T(t_0)$ be the 3D translation and $R(t_0)$, the rotation, needed for shifting the reference frame w.r.t. the movement commencing at time t_0. $T(t_0) = -[x_o(t_0), y_o(t_0), z_o(t_0)]^T$, where x_o and y_o are as shown in Figure 3.7(a), and z_o is the height of the toes of the subject in the world-centric frame. The rotation matrix $R(t_0)$ defines an anti-clockwise rotation by θ (see Figure 3.7(a)).

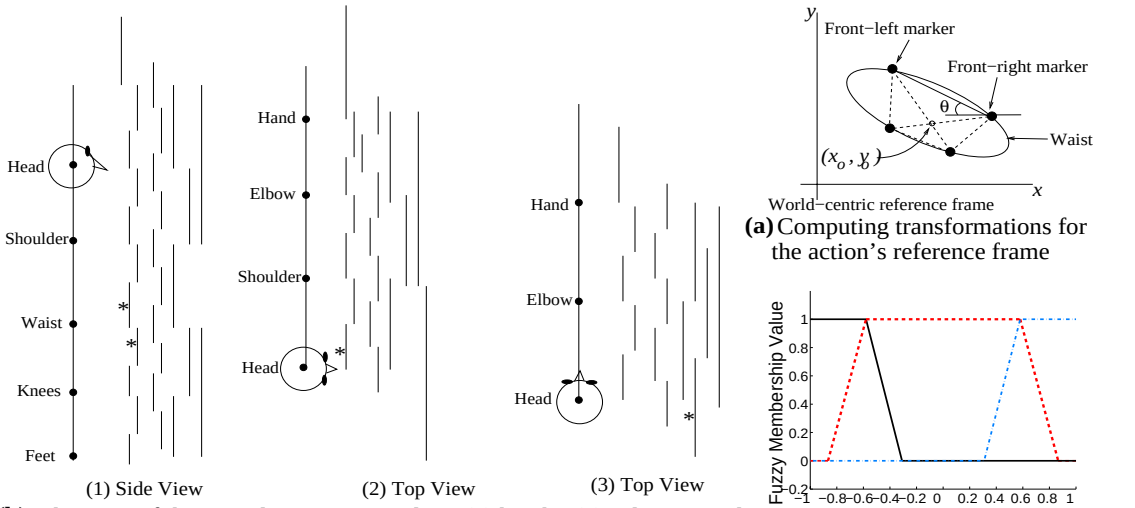
Let $\mathbf{x}(t)$ be the 3D coordinates of a body part as given by motion capture,

where $t \in [t_0, t_1]$. These would be in world-centric coordinates. The analysis is done on the transformed coordinates $\tilde{\mathbf{x}}(t) = R(t_0)[\mathbf{x}(t) + T(t_0)]$.

3.6.4.2 Location of Target and Direction of Movement

Location: A 3D orthogonal coordinate system is employed for representing the target's location. This could simply be the target's 3D Cartesian coordinates in the movement's reference frame. However, comparing the similarity/dissimilarity of the target locations of the movements would be difficult. Instead, we quantize the space around the subject in terms of his/her morphology. For example, the dimension along the height axis is quantized into regions such as "at feet level", "below knee level", "at knee level", etc. The reasoning is that, in the absence of external reference points obtained from the environment, humans reference their immediate neighborhood in terms of their own morphology [79]. The regions overlap and are of different sizes. Examples of the volumes obtained are: in front of the chest, in front of the left half of the chest, etc. See Figure 3.7(b) for a schematic of the spatial quantization.

Direction: Similar to spatial location, the movement direction is also described using labels. Let $\mathbf{d}(t) = \frac{\tilde{\mathbf{x}}(t+1) - \tilde{\mathbf{x}}(t)}{\|\tilde{\mathbf{x}}(t+1) - \tilde{\mathbf{x}}(t)\|}$ be the unit direction vector of movement at time t. The x component of $\mathbf{d}(t)$ is divided into forward, negligible and backward motion, the y component into leftward, negligible and rightward motion, and the z component into upward, negligible and downward motion. Therefore, each compo-



(a) Computing transformations for the action's reference frame

(b) Schematic of the spatial quantization along (1) height, (2) sideways and (3) front/back directions. The line segments indicate the bins. In case of sideways and front/back directions, the shown bins are reflected about the head to allow for locations on the right and at the back. Bins marked with a '*' define the resting position for the hands. Movements with targets located in these bins are considered to be irrelevant.

(c) Quantization of of each component of the unit direction vector into 3 bins. The 3 plots (solid, dashed and dot−dashed) show the membership values assigned to each bin.

Figure 3.7: (a) Computing the movement's reference frame, (b) Spatial Quantization, and (c) Direction quantization.

nent of the unit direction vector is quantized into three bins having angular width of 120° - shown in Figure 3.7(c). Let $\hat{\mathbf{d}}_x(t)$ denote a 3×1 vector quantifying the membership values of the x component of the direction vector in the 3 bins. The membership values vary continuously from 0 to 1. Similarly, $\hat{\mathbf{d}}_y(t)$ and $\hat{\mathbf{d}}_z(t)$ are defined for the y and z components respectively. The complete quantization is denoted by $\hat{\mathbf{d}}(t) = [\hat{\mathbf{d}}_x(t)\ \hat{\mathbf{d}}_y(t)\ \hat{\mathbf{d}}_z(t)]$.

3.6.5 Experimental Results

The proposed approach was tested with several capture sequences of reach and strike movements. These included cases in which a subject assembles and uses a vacuum-cleaner, moves around objects, climbs a ladder, etc. For the strike

movements, the subjects pretended as if boxing - they stepped around, dodged and executed combinations of punches, jabs, hooks, etc. The duration of the sequences varied from 3 sec. to approximately 40 sec. The data used for training and testing was obtained from different subjects so as to observe the generalization ability of the approach. The ground truth for each sequence was manually observed. Out of 55 instances of reach movements, 44 (80%) were detected correctly and there were 2 false detections. Some of the reach movements were missed due insubstantial movement of the hands. There were also cases during the vacuum-cleaner assembly in which it was not clear if the movements were ballistic - these were still considered as reaches in the ground truth. Out of 78 instances of strike movements, 71 (91%) were detected correctly and there were 6 false detections. The 6 false strike detections were for cases when the subject made rapid hand movement before executing a "hook". Figures 3.4 and 3.8 show the labels generated for some instances of striking and reaching. For Figure 3.8, the movements were: (a) Subject takes a step forward and reaches out forward with right hand near knee level, (b) Subject turns around and takes a couple of steps to reach out behind with right hand, and (c) Subject reaches for the floor and then above the head. As is illustrated in the figures, the target labels generated by the proposed approach are coherent.

The analysis of motion capture data indicates that the ballistic movement model enables generalization over the subjects, and accurate recognition of reach and strike movements when the hand's location is available. The next section develops this concept into a video-based recognition system. The unavailability of body-part trajectories, noise present in the video and ambiguity in pose estimation

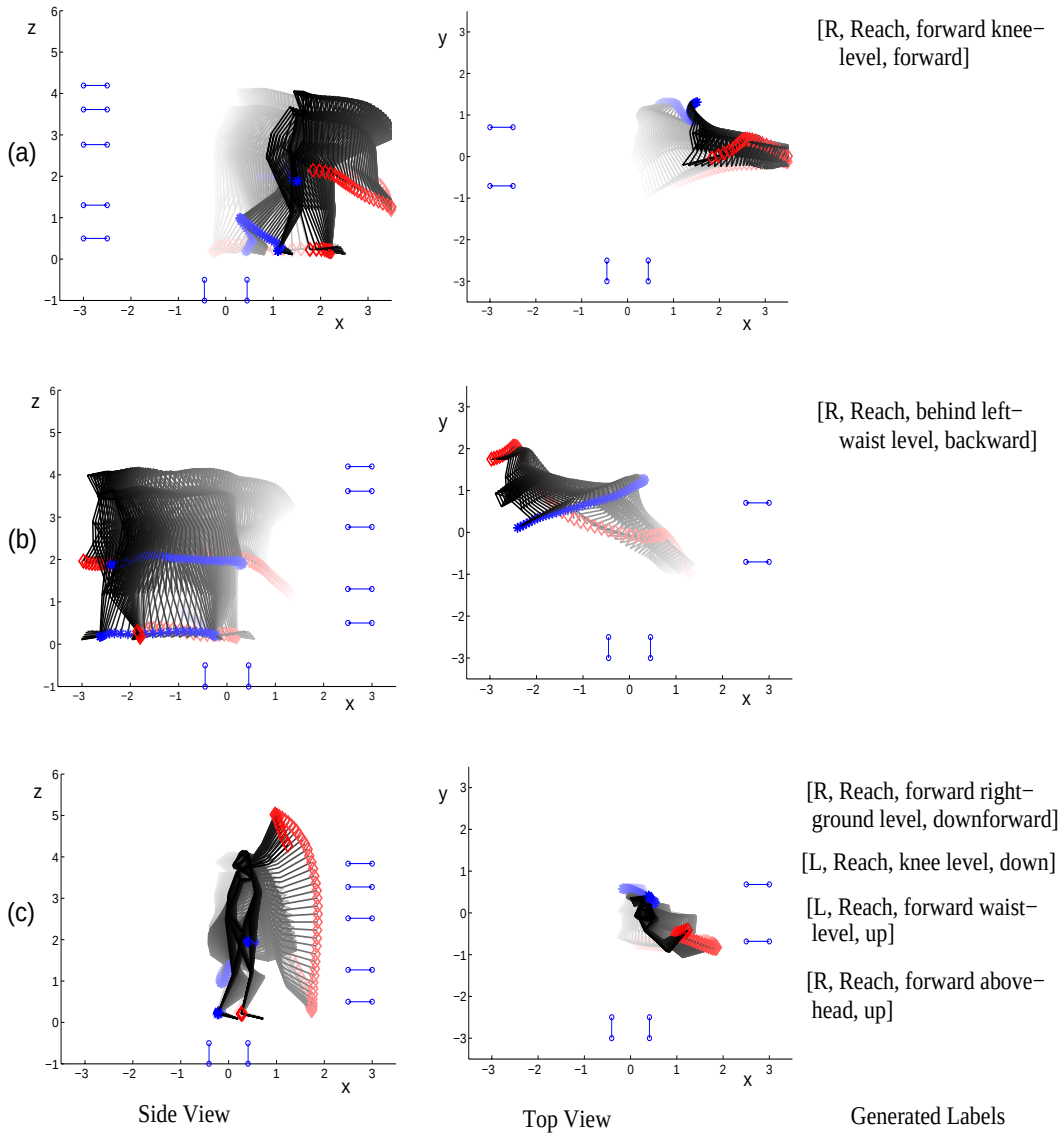


Figure 3.8: Examples of the labels generated - shown in the sequence in which they were output.

make visual recognition challenging. These are addressed by employing state-of-the-art machine learning techniques for modelling the statistics of low-level image and motion features for the recognition.

Chapter 4

Video-Based Analysis of Ballistic Hand Movements

Tracking the hands and the pose is one of the most challenging aspects of human action recognition. Is it possible to analyze the movement dynamics and perform recognition without pose-tracking? We explore this question by modelling the statistics of low-level image and motion features. Section 4.1 describes the motion and image features used to implicitly represent the hand's velocity. Next, a Dynamic Programming algorithm is presented that efficiently computes the optimal segmentation of a sequence into ballistic movements. The segments are classified as reaches and strikes based on the statistics of the motion features. Finally, movement labels are inferred based on the person's pose at the start and end of movement.

4.1 Representing the Hand's Velocity

Due to the correlation in the body-parts' velocities during ballistic movements, the hand's velocity can be implicitly represented with low-level motion features computed over the entire figure of the person. This does not require the hands and the arms to be isolated/segmented from the rest of the body. The term "low-level" refers to features that capture the gross motion flow of the movement without explicitly tracking the body parts. This enables the system to perform recognition even when the hands and arms cannot be accurately localized due to occlusion, edge

clutter and rapid movement. In our study, the motion features consist of optical flow, silhouette deformation and frame differences.

4.1.1 Optical flow

We employ a phase-based optical flow approach proposed by Gautama et al. [29]. Background subtraction is used to obtain the set of optical flow vectors located on the subject's silhouette [73]. Let F_t denote the set of optical flow vectors obtained at time t. The utility of optical flow is illustrated using two experiments:

(a) Are the flow vectors mutually consistent, i.e. pointing in the same direction?

(b) Do the flow vectors have high correlation on the direction of hand's movement?

4.1.1.1 Self-consistency of Optical Flow Within a Movement

A video clip consisting of 12 reach movements performed by a subject was analyzed. Let $[t_s^i, t_e^i]$ denote the time interval of the i^{th} movement. The set of flow vectors obtained for the i^{th} movement would be $\bigcup_{t=t_s^i}^{t_e^i} F_t$. The self-consistency of the optical flow during a movement is measured by the dot product of the flow vectors w.r.t. the mean flow vector for the movement. Figure 4.1(a) shows the histogram of the values of self-consistency obtained for the movements. It indicates that most of the optical flow vectors point in the same direction as the mean flow vector, highlighting the self-consistency of the flow. Flow vectors whose dot product with the mean flow is greater than 0.5 are considered to be relevant for measuring the movement's dynamics; they constitute the *significant* optical flow, $\mathcal{F}_i$ for the i^{th}

movement.

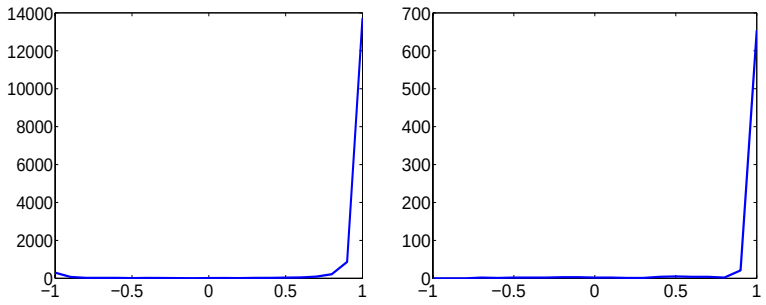
4.1.1.2 Consistency of the Hands' Direction of Movement with the Optical Flow

Next, we measure the consistency between the direction of the 2D projective velocity of the subjects' hands during reach movements and the optical flow computed over his/her silhouette. A video sequence of several reach movements was collected and the subject's hands' centroids in the image frames were hand-labelled. Let $\mathbf{v}_t$ denote the displacement vector of a hand at time t computed using 1^{st}-order differences. It's consistency with the optical flow is defined to be it's normalized dot-product with 5 Nearest-Neighbor *significant* optical flow vectors, formulated as

$$\mathbf{v}_t \odot \mathcal{F}_i = \text{mean\_5NN}\left(\left\{\frac{\mathbf{v}_t}{\|\mathbf{v}_t\|}\cdot\frac{\mathbf{f}}{\|\mathbf{f}\|}\,\middle|\, t \in [t_s^i, t_e^i] \wedge \mathbf{f} \in \mathcal{F}_i\right\}\right)$$

As the hands' size in the image frame was typically 10×10, displacement vectors with $\|\mathbf{v}\| \leq 5$ were ignored. Figure 4.1(b) shows a histogram of the values of the consistency measure observed for the movement sequence. It indicates that a majority of the hands' displacement vectors have high consistency with the optical flow.

The flow at time t is represented by the mean vector of F_t, i.e. by $\tilde{\mathbf{f}}_t = \frac{\sum_{\mathbf{f}\in F_t}\mathbf{f}}{|F_t|}$. The magnitude of the optical flow vectors is noisy due to the rapidity of the movements and the small visual area occupied by arms and hands. To provide robustness, the flow magnitude is represented by the $\min(\|\tilde{\mathbf{f}}_t\|)$, $\text{mean}(\|\tilde{\mathbf{f}}_t\|)$, $\text{median}(\|\tilde{\mathbf{f}}_t\|)$ and $\max(\|\tilde{\mathbf{f}}_t\|)$ within small temporal windows. In our experiments, 5 window sizes were



(a) (b)

Figure 4.1: (a) Histogram of the dot product of optical flow vectors with the mean optical flow vector, (b) Histogram of the dot product of instantaneous displacement vector of the hand with 5-NN optical flow vectors.

used, of lengths 6 to 10. This results in a 20 dimensional feature vector, $\tilde{\mathbf{\Phi}}_{\text{OptFlow}}(t)$, representing the magnitude of the optical flow at time t. Note that by eliminating the directional information, the features are designed to be invariant to the direction of movement.

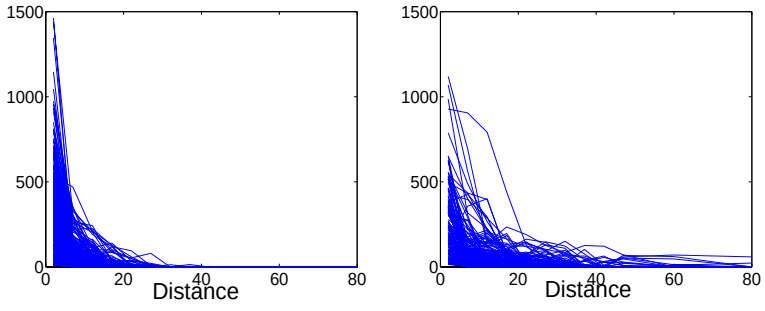
4.1.2 Silhouette Deformation

The subject's silhouette in each frame is computed using background subtraction followed by contour extraction [73]. A Distance transform $D_t(\mathbf{x})$ is computed on the image plane for the silhouette at each time instant t. The deformation of a silhouette at time t is measured by the Chamfer distance of the points on the silhouette w.r.t. $D_{t-1}(.)$. Let $\{\mathbf{p}_1^t, \ldots, \mathbf{p}_N^t\}$ be the points on the silhouette at time t. Let S_t be the set of Chamfer distances at these points, $S_t = \{D_{t-1}(\mathbf{p}_i^t)\}_{i=1}^N$. It is summarized using four measures: $\min(S_t)$, $\text{mean}(S_t)$, $\text{median}(S_t)$ and $\max(S_t)$. These measures are averaged (mean and median) over various time windows to achieve robustness to noise. A 20 dimensional feature vector is created for each time instant, denoted

by $\tilde{\mathbf{\Phi}}_{\text{SilDef}}(t)$.

4.1.3 Pixel-wise Frame Differences

Motion-history images and pixel-wise differences have been extensively used to represent motion [11, 89]. Let $I_t(\mathbf{x})$ denote the image at time t. The difference image is defined as $\delta I_t(\mathbf{x}) = \lfloor (I_t(\mathbf{x}) - I_{t-1}(\mathbf{x})) > \Delta_{\text{ID}} \rfloor$. The threshold Δ_{ID} depends upon the noise characteristics of the video and is fixed at 0.1. A distance map $D_t^{\delta}(\mathbf{x})$ is constructed from $\delta I_t(.)$. Let ID_t be the set of Chamfer distances of active pixels in $\delta I_t(.)$ w.r.t. $D_{t-1}^{\delta}(.)$. It is defined as $\text{ID}_t = \{D_{t-1}^{\delta}(\mathbf{x}) | \delta I_t(\mathbf{x}) = 1\}$. A histogram is constructed at each time instant from the members of ID_t; it is quantized so as to reduce the effects of noise and outliers. Figure 4.2 shows line-plots of the histograms obtained during mid-flights of some reach and strike movements. They indicate that strike movements have higher frequency of large displacements. The histograms represent the velocity as a 12 dimensional feature vector, $\tilde{\mathbf{\Phi}}_{\text{FrmDiff}}(t)$, for each time instant.



(a) (b)

Figure 4.2: Histograms of ID_t computed during mid-flight for (a) reach and (b) strike movements. The plots indicate that strike movements have higher frequency of large displacements.

4.1.4 Summary of Velocity Features

The silhouette deformation and pixel difference features involve statistics of the displacement magnitudes - Chamfer Distance. Therefore, they are robust to changes in movement direction. The magnitude of the velocity of the hand is represented by

$$\tilde{\mathbf{\Phi}}(t) = \left\langle \tilde{\mathbf{\Phi}}_{\mathrm{OptFlow}}(t), \tilde{\mathbf{\Phi}}_{\mathrm{SilDef}}(t), \tilde{\mathbf{\Phi}}_{\mathrm{FrmDiff}}(t) \right\rangle.$$

The acceleration and deceleration impulses are represented by including past and future velocity magnitudes to obtain

$$\mathbf{\Phi}(t) = \left\langle \tilde{\mathbf{\Phi}}(t-2\Delta t),\ \tilde{\mathbf{\Phi}}(t-\Delta t),\ \tilde{\mathbf{\Phi}}(t),\ \tilde{\mathbf{\Phi}}(t+\Delta t),\ \tilde{\mathbf{\Phi}}(t+2\Delta t) \right\rangle$$

$\mathbf{\Phi}(t)$ depends only upon the magnitude of the motion, and thus, is robust to variation in direction of movement and camera-view. By including past and future velocity information, it implicitly represents statistics of accelerating and decelerating impulses. This enables it to encode the ballistic dynamics of the hand for classification into reaches and strikes - described in Section 4.3. The velocity observations, $O_{\mathrm{v}}(t)$, consist of $\mathbf{\Phi}(t)$ to encode velocity magnitude and $\tilde{\mathbf{f}}(t)$ to represent the direction of motion.

4.2 Temporal Segmentation into Ballistic Movements

Recalling from Section 3.3, sequences are segmented into ballistic movements by fitting the dynamical model to subsequences of motion observations and noting the segmentation with maximum likelihood. We describe the manner of fitting the ballistic dynamics to subsequences, and then the Dynamic Programming algorithm

for efficiently computing the optimal segmentation.

It can be shown that in the MJM model, when initial and final velocity and acceleration are zero, the hand follows the following trajectory

$$\mathbf{y}(t) = \begin{bmatrix} y_1(t_s) + (y_1(t_s) - y_1(t_e))(15\tau^4 - 6\tau^5 - 10\tau^3) \\ y_2(t_s) + (y_2(t_s) - y_2(t_e))(15\tau^4 - 6\tau^5 - 10\tau^3) \end{bmatrix} \tag{4.1}$$

where $\mathbf{y}(t_s) = [y_1(t_s), y_2(t_s)]$ is the initial position, $\mathbf{y}(t_e) = [y_1(t_e), y_2(t_e)]$ is the ending position, and $\tau = \frac{t-t_s}{t_e-t_s}$ is the time scale. It is easy to see that the trajectory is a straight line. A number of psychological studies have noted this to be a good approximation of the path followed by the hand during reach movements e.g., [79, 28, 85], etc. We employ it for approximating the path followed by the hands during ballistic movements. As will be shown in the experiments, this forms a good assumption given the high acceleration and deceleration involved, and the relatively short duration of the movements.

Consider the i^{th} segment of duration $[t_s^i, t_e^i]$. Let the direction of movement of the hand be θ_i - this parameterizes the dynamics B_i. The likelihood of θ_i's fit to $O_{\text{v}}(t_s^i) \ldots O_{\text{v}}(t_e^i)$ is defined through potential functions on the weighted difference between the optical flow vectors and θ_i direction:

$$p(O_{\text{v}}(t_s^i) \ldots O_{\text{v}}(t_e^i)|B_i) = \prod_{t=t_s^i}^{t_e^i} \prod_{\mathbf{f} \in F_t} \exp - [\|\mathbf{f}\| - \mathbf{f} \cdot \hat{\mathbf{n}}(\theta_i)] \tag{4.2}$$

where $\hat{\mathbf{n}}(\theta) = \cos\theta\hat{i} + \sin\theta\hat{j}$. Taking a logarithm and differentiating with respect to θ_i, the optimal value of fit is obtained for

$$\sum_{t=t_s^i}^{t_e^i} \sum_{\mathbf{f} \in F_t} (f_1 \sin\theta_i - f_2 \cos\theta_i) = 0 \tag{4.3}$$

Therefore, the optimal value of $p(O_{\mathrm{v}}(t_s^i)\ldots O_{\mathrm{v}}(t_e^i)|B_i)$ is

$$p(O_{\mathrm{v}}(t_s^i)\ldots O_{\mathrm{v}}(t_e^i)|B_i^*) = \exp\left(\left\|\sum_{t=t_s^i}^{t_e^i}\sum_{\mathbf{f}\in F_t}\mathbf{f}\right\| - \sum_{t=t_s^i}^{t_e^i}\sum_{\mathbf{f}\in F_t}\|\mathbf{f}\|\right) \tag{4.4}$$

From Eq.(3.8), we have the probability of the segmentation of a sequence into ballistic segments $B_1\ldots B_n$ as

$$p(B_1\ldots B_n|O_{\mathrm{v}}) = \exp\left(\sum_{i=1}^{n}\left\|\sum_{t=t_s^i}^{t_e^i}\sum_{\mathbf{f}\in F_t}\mathbf{f}\right\| - \sum_{i=1}^{n}\sum_{t=t_s^i}^{t_e^i}\sum_{\mathbf{f}\in F_t}\|\mathbf{f}\|\right) \tag{4.5}$$

Notice that $\sum_{i=1}^{n}\sum_{t=t_s}^{t_e}\sum_{\mathbf{f}\in F_t}\|\mathbf{f}\|$ is a constant for the sequence, independent of the segmentation. Therefore, the optimality of the segmentation of a sequence $[0,T]$ into partition $\chi = \langle\chi_0 = 0, \chi_1, \ldots, \chi_n = T\rangle$ is given by

$$\sum_{i=1}^{n}\Psi(t_s^i, t_e^i)$$

where

$$\Psi(t_s^i, t_e^i) = \left\|\sum_{t=t_s^i}^{t_e^i}\sum_{\mathbf{f}\in F_t}\mathbf{f}\right\| \tag{4.6}$$

Let the minimum duration of a ballistic movement be $T_{\min}$ and the maximum duration be $T_{\max}$. In our experiments, $T_{\min} = 5$ frames (0.25 sec.) and $T_{\max} = 30$ frames (2 sec.) Algorithm 1 describes an $\mathrm{O}(n)$ algorithm for computing the optimal segmentation. Figures 2.1, 3.2, 4.5 and 4.6 show examples segmentations computed for reach and strike sequences. Quantitative results are presented in Section 4.5.

4.3 Reach vs. Strike Classification

Ballistic movement segments are classified into reaches and strikes by modelling the statistics of the motion features. A classifier based on boosting was trained

Algorithm 1 Temporal_Segmentation

Procedure

Set $\Psi(t_i, t_j) = 0 \; \forall t_i, t_j < 0$ /*Boundary condition*/

for $t_s = 0 \ldots T$ **do**

$\hat{\mathbf{f}}(t_s, t_s) = \sum_{\mathbf{f} \in F_{t_s}} \mathbf{f}$

$$\chi(t_s) = \begin{cases} -1 & t_s = 0 \\ \arg\max_{t=t_s-T_{\min}}^{t_s-T_{\max}} \hat{\Psi}(t, t_s) & t_s > 0 \end{cases}$$

$$\Psi_{\text{best}}(t_s) = \begin{cases} 0 & t_s = 0 \\ \max_{t=t_s-T_{\min}}^{t_s-T_{\max}} \hat{\Psi}(t, t_s) & t_s > 0 \end{cases}$$

for $t_e = t_s + 1 \ldots t_s + T_{\max}$ **do**

$\hat{\mathbf{f}}(t_s, t_e) = \hat{\mathbf{f}}(t_s, t_e - 1) + \sum_{\mathbf{f} \in F_{t_e}} \mathbf{f}$

$\Psi(t_s, t_e) = \|\hat{\mathbf{f}}(t_s, t_e)\|$

$\hat{\Psi}(t_s, t_e) = \Psi(t_s, t_e) + \Psi_{\text{best}}(t_s)$

end for

end for

$\chi(T) = \arg\max_{t=T-T_{\max}}^{T-T_{\min}} \hat{\Psi}(t, T)$

$\Psi_{\text{best}}(T) = \max_{t=T-T_{\max}}^{T-T_{\min}} \hat{\Psi}(t, T)$

$\chi^* = T$ /*Recursively backtrack to get optimal segmentation*/

$t = \chi(T)$

while $t \neq -1$ **do**

$\chi^* = t \oplus \chi^*$ /*Concatenation operator*/

$t = \chi(t)$

end while

$\Psi^* = \Psi_{\text{best}}(T)$

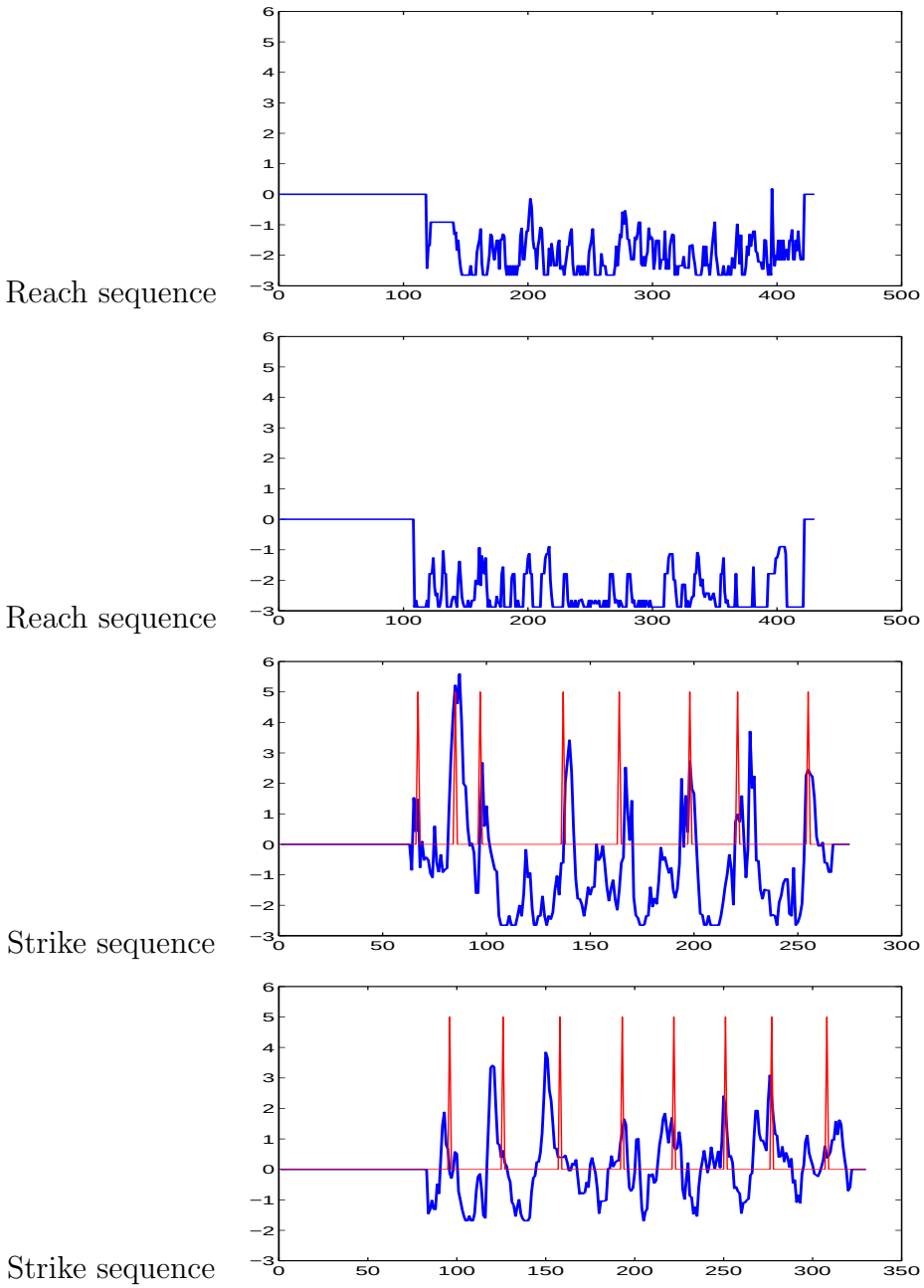


Figure 4.3: Confidence values of strike detection for two reach movement sequences and two strike sequences. The ground-truth timing of strike movements are marked with a red impulse function.

to distinguish between instants of reach and strike movements [1]. The data vectors were constructed from the motion features described in Section 4.1. The training samples consisted of feature vectors computed at mid-flight during reach and strike

movements. Let $h(.)$ denote the trained classifier, whose output is 1 for strike movement dynamics and -1 otherwise. The confidence for detection of strike at time t is defined as $h(\mathbf{\Phi}(t))$ - higher the value more the likelihood of strike movement. Figure 4.3 shows examples of the confidence values as a function of time for two sequences consisting entirely of reach movements (no striking or throwing), and two sequences in which people threw objects and punched around. The ground-truth time of the strike movements are marked as impulses in a red plot overlayed on them. In the plots of strike sequences, $h(.)$ has peaks corresponding to strike movements, indicating that the classifier is able to distinguish between reach and strike movement dynamics. Quantitative results are presented in Section 4.5.

4.4 Position Features and Label Inference

The subject's pose, $O_{\mathrm{p}}(t)$, is represented by the subject's silhouette and the head's gaze-direction. Shape-Context, proposed by Belongie et al. [9], is used to represent the subject's silhouette. The subject's gaze-direction w.r.t. the camera is represented by a 4D vector of confidences in four gaze-directions: left, right, facing the camera and facing away from the camera. Gaze-detection has been extensively studied as part of pose-invariant head detection, e.g., [39]. A simple gaze-detector based on Haar-like features is used to determine the head's gaze-direction. The hand's position is estimated using skin detection and motion features [73]. Figure 4.4 shows some examples of the silhouette, head's pose, and hand's target location detected for some reach and strike movements.

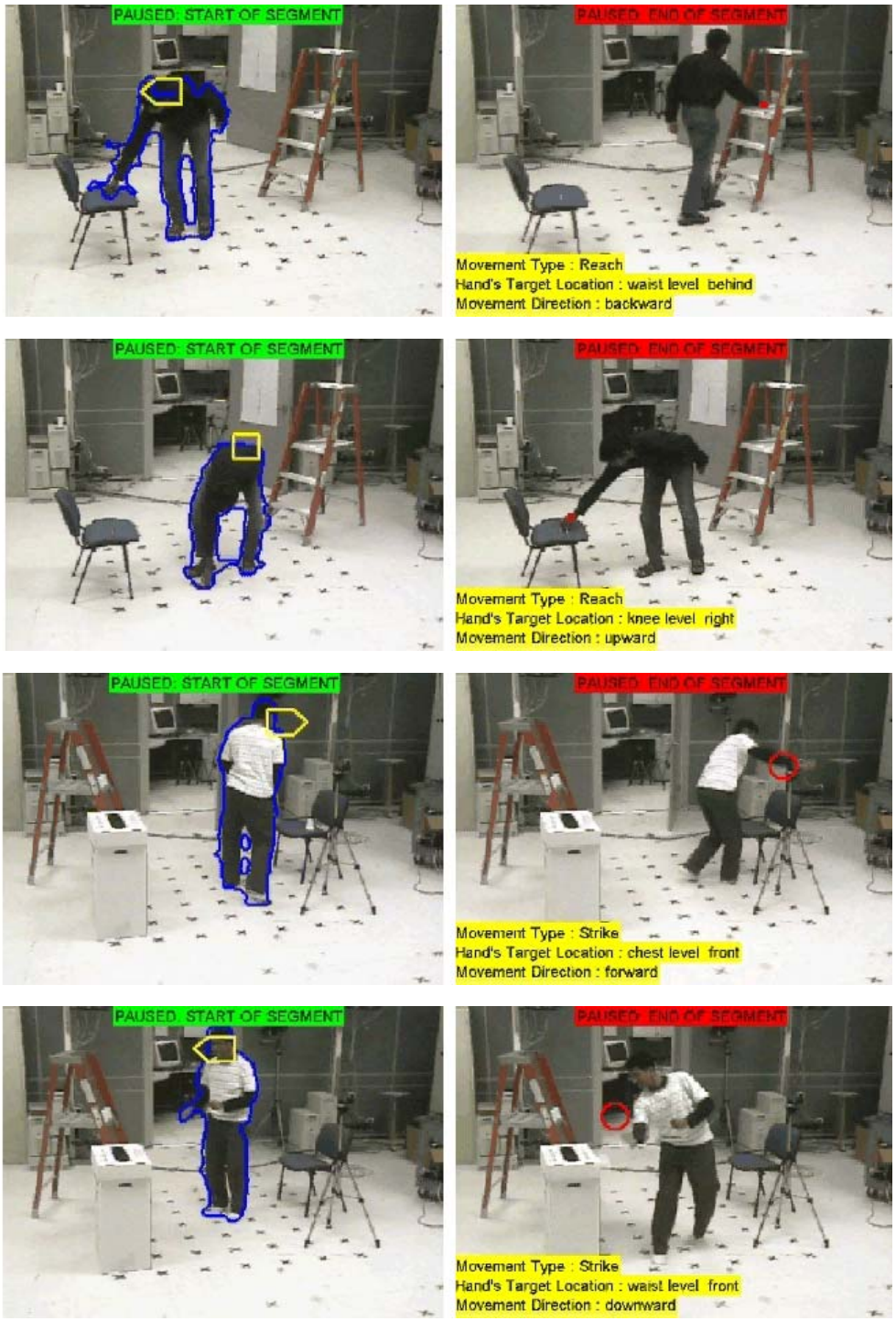


Figure 4.4: Examples of the person silhouette and gaze-direction computed at the start of ballistic movements, and the hand's target location estimated using skin detection and motion.

The silhouette and the head's gaze-direction provide spatial context for labelling the hand's target location. For example, the position of the hand relative to the principle axis of the silhouette depends upon the height of the hand in world coordinates. Similarly, the head's gaze direction determines labels such as front, left, behind, etc. Figures 2.1, 4.5, 4.6 illustrate the labels computed by the approach. Quantitative results are presented in the next section.

4.5 Experimental Results

A database of movement sequences was collected to test the approach: 7 reach movement sequences were collected depicting 67 reach instances performed by 6 subjects. A number of small objects such as pens, clips, etc. were placed on surfaces of varying heights in the scene. The subjects were asked to pick up and place the objects on random surfaces of their choice including the floor. They were asked to confine their movements within an area of 9×9 feet. No restriction was imposed on the manner of movements - the subject stepped around, bent, used either of their hands, etc. Based on their own volition, subjects performed movements in rapid succession as well as with pauses. The segmentation of continuous sequences into ballistic movement segments was performed automatically. Movement instances in which the hands were occluded were ignored.

In a similar manner, we recorded 10 strike sequences depicting 68 instances of striking and throwing performed by 4 subjects. The subjects were asked to strike and throw objects kept at various heights varying from the ground to waist-level. No

restriction was imposed on the manner of the strikes - subjects punched, slammed down and slapped (forehand and backhand) the objects. The subjects struck and threw with all their might - one subject almost broke a garbage bin while slamming down on it!

The subjects consisted of 5 males and 1 female - the subjects' morphologies vary considerably. The video resolution was 320×240, at 15 frames per second. The subjects' heights in the image-frames were $\approx 180 \pm 40$ pixel units.

The data-set is challenging as many movements are executed in rapid succession and at high speeds. The limbs are frequently inside the subject's silhouette, making pose-estimation difficult. There is significant motion blur during mid-flight. Please see supplementary videos. Table 4.1 shows the recognition results for the reach and strike movements.

Segmentation results are shown in Row 2 of Table 4.1. Very few movements were missed by the segmentation. The error in the boundary of the segments was in the range ± 3 frames (0.2 sec). A likely reason for this error is that the hand's velocity during the first few and last few frames of a movement segment is very low. Low level motion features are inadequate for such fine differentiation.

Reach vs. strike classification results are shown in Rows 3 and 4 of Table 4.1. The accuracy is high, the error rates being approximately 6%. In 2 of the cases in which strike movements were misclassified as reaches, the strike movement's duration was very small (2 to 3 frames). Due to the noise present in images and the subject's silhouette, it is difficult to reliably extract motion features for movements of such short duration.

	Ground truth classes	
	Reaches	Strikes
1. Total number of instances (ground-truth)	67	68
2. Num. correctly segmented	64	68
(percentage)	96%	100%
3. Num. classified as reaches	60	4
(percentage)	**90%**	6%
4. Num. classified as strikes	4	64
(percentage)	6%	**94%**
5. Correct reach/strike classifications & labelling of movement's direction and target location	56	59
	84%	**87%**

Table 4.1: Video-based movement recognition results

Target location & Movement direction results are shown in Row 5 of Table 4.1. The total number of reach movements that were correctly detected, classified and qualitatively labelled was 56 (84%). 2 of the target labelling errors were due to incorrect estimation of the hand's position at the end of the movement.

The total number of strikes correctly detected, classified and qualitatively labelled was 59 (87%). There are two reasons for the errors in labelling strikes:
(1) At very high speeds, the hand's image is blurred. For strike sequences with pronounced blurring of the hand, the target's position at the time of highest speed

may not be detected, resulting in incorrect labels.

(2) The optical flow computation is unreliable for rapid movements of very short duration. This resulted in erroneous labels for the direction of movement for some instances of strikes.

Figure 4.5: Labels generated for three reach movements. To save space, every third frame of the sequences are shown.

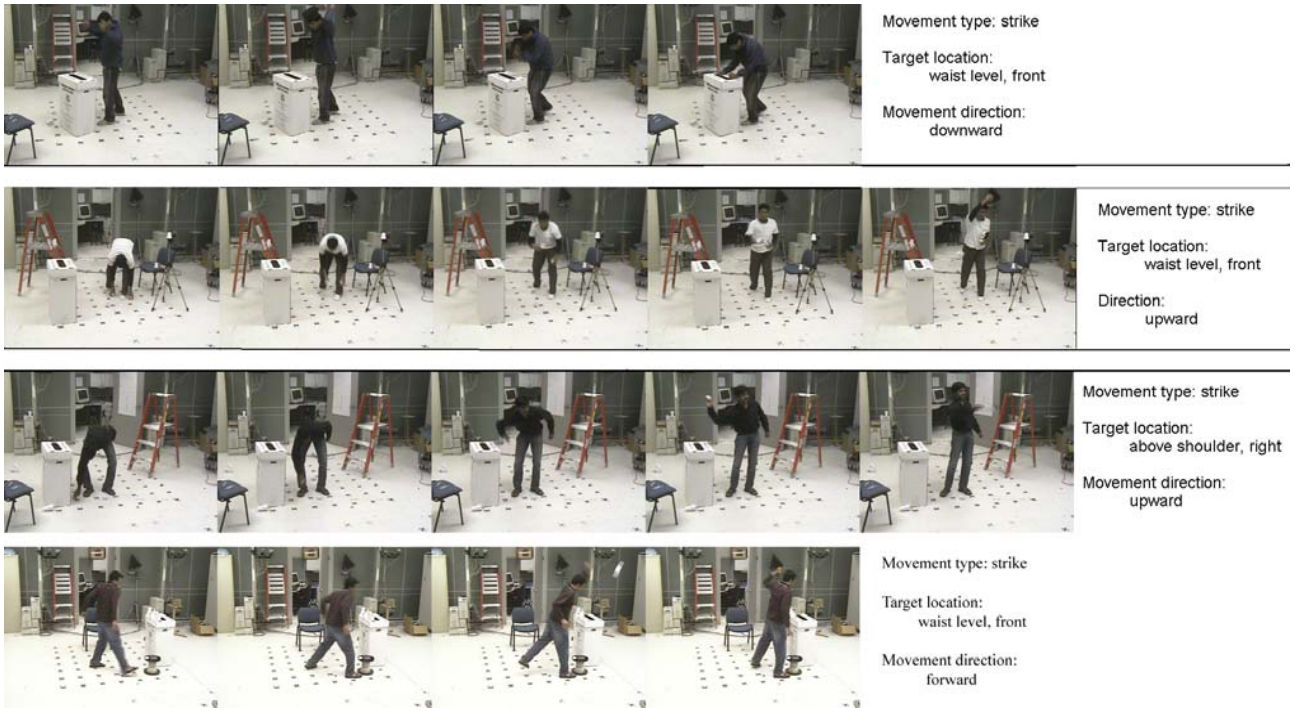


Figure 4.6: Labels generated for four strike movements. To save space, every third frame of the sequences are shown. (Best viewed in color.)

Chapter 5

Edge Continuity for Contour Matching

5.1 Introduction

Edge continuity has been studied in computer vision to understand the manner in which humans organize and group visual structures. Kanizsa's experiments with subjective contours was one of the earliest papers on this subject [42]. Later, several computational approaches were proposed to model edge continuity, showing interesting resemblance to human perception, e.g., [74, 58, 34, 90]. Parallel to this, there has been research on human pose estimation and detection using contour matching. Recognition is performed by matching model contours with image edges. Edge clutter present in natural images is one of the principle challenges faced during contour matching. We explore edge continuity models for improving recognition, and apply it to human pose estimation and gesture recognition. This chapter presents an edge affinity model that extends previous approaches by including color statistics in the neighborhood of edges. The model is employed for improving pose estimation. Results of this work were reported in [62]. Chapter 6 presents a Markov Random Field (MRF) extension, and applies it to human detection.

5.1.1 Studies on Contour Matching

Contour matching is used extensively in computer vision for human pose detection and recognition tasks. When applied for action or gesture recognition, it is used to compute pose observation likelihoods, which are then modelled using Hidden Markov Models (HMMs) [24], Markov Chain Monte Carlo (MCMC) [47], etc. Contour matching has also been used for object detection e.g. [30, 54], etc. There are three stages to contour matching:

1. Edge features of the objects in the images are detected.

2. A pose-contour is imposed on the image for matching.

3. The score for the match is generated by computing distance between the image's edge features and the imposed pose-contour.

Many studies - including ours - use gradient-based operators such as Canny edge-detector and Gaussian derivatives for detecting edge features. Reliably detecting object boundaries in general illumination conditions is difficult. Recent research on boundary detection has focussed on using region segmentation as a pre-processing step for generating "super-pixels" - relatively small groups of pixels that have homogenous features and are highly likely to belong to the same object. Boundaries of the super-pixels are used for matching object boundaries. For example, Mori et.al. use normalized-cuts (n-cuts) to obtain super-pixels and then analyze their configurations to detect baseball players [56]. Sharon et.al. use a multigrid approach for obtaining segment boundaries [75].

The pose-contour to be matched with the test image could either be collected during training or generated using a model. Whole-body contours have been used for human pose-matching in [57, 31, 84, 55, 24], etc. Zhang et al. use a Bayes-nets based articulated model for pedestrian detection [98]. Ronfard et al. follow a bottom-up part-based approach to detecting people [71]. They train Support Vector Machines (SVMs) on gradients of limbs obtained from training images. In the present study, the pose-contours correspond to the whole body of the subject and are collected during a training phase.

Chamfer distance is a popular method for measuring the goodness of the match between edge sets. The distance for each contour point from the nearest image edge is computed. The sum of these distances indicates the goodness of the match - lower the integral, better the match. Rosin and West presented a continuous form of chamfer distance which includes the saliency of the edges in the matching [61]. Their method avoids setting threshold on the gradient magnitudes, generally difficult issue. Butt and Maragos presented an efficient approach for computing chamfer distance while minimizing errors due to discretization [17]. Toyama and Blake use sets of exemplar contours and chamfer distance for tracking pedestrians and mouth movements [84]. Mori and Malik introduced the Shape Context technique for matching human pose contours [55]. Olson and Huttenlocher used the Hausdorff distance for object recognition [57]. Leibe et.al. present a study comparing contour-based and appearance-based object recognition in [48].

5.1.2 Pose Matching in Cluttered Images

Images of people in natural scenes have significant edge clutter present in the background in addition to the subject's figure. Ideally, these background edges should be ignored when matching pose-contours. However, reliable background suppression in natural images in the presence of camera and subject motion is difficult. There are three general ways of handling this:

5.1.2.1 Asymmetric Approach

Not perform the difficult task of background subtraction but rather compromise with asymmetric matching, which only measures how well a model pose-contour matches with the image's gradients. It does not verify whether these matching gradients form a coherent object. Current contour matching schemes either follow this asymmetric approach or assume background subtracted images, e.g. [24, 30, 54, 57, 17, 84, 55, 98]. Predictably, this leads to problems as a contour can match well with a subset of the edges of an object and ignore the rest of it. Consider the case shown in Fig. 5.1. Fig.s 5.1(a) and (e) show an image and the edges of the subject. Fig.s 5.1(c) and (g) show two pose-contours in the database extracted from training images shown in Fig.s 5.1(b) and (f) respectively. Clearly, the contour in Fig. 5.1(c) is the correct pose. However, when the poses are matched with the image (Fig.s 5.1(d) and (h)) using Chamfer matching, the wrong pose obtains a better score. The reason is that it has smaller extent at the arms, which - due to articulation - are the zones of highest errors in matching. Normalizing the error

w.r.t. the length of the boundary does not ameliorate the situation.

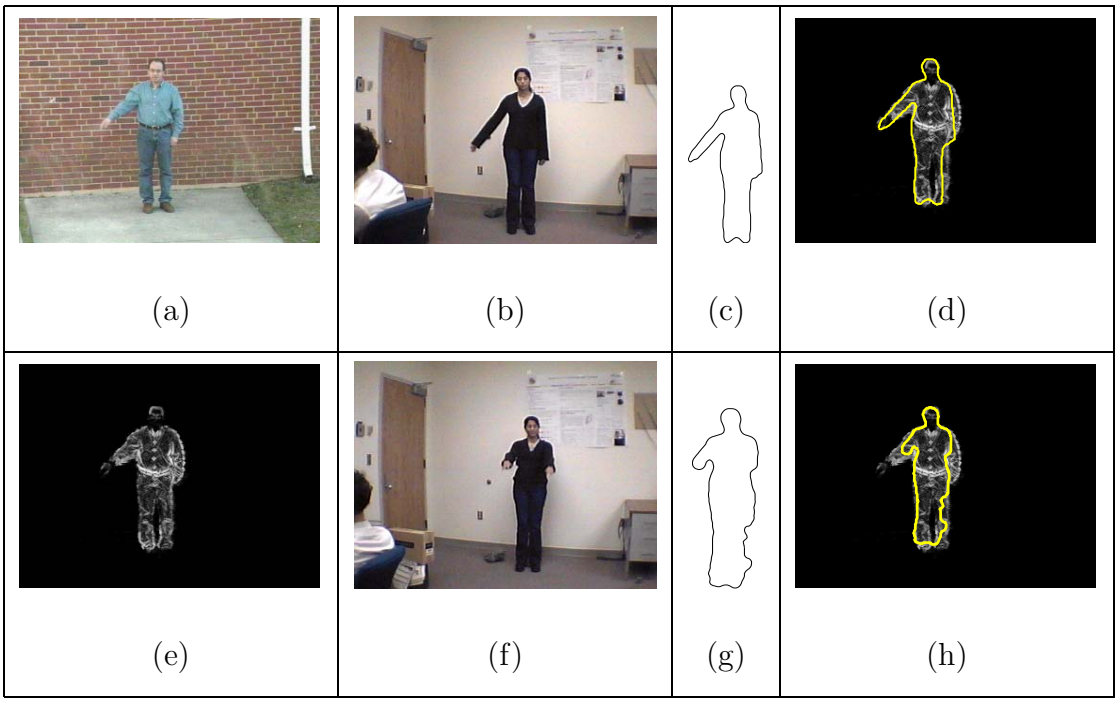


Figure 5.1: (a,e) The test image and the subject's edges. (b,c,d) Training image showing the correct pose, the pose extracted from it, and the gradient map of the test image with the pose overlayed. (f-h) Similar to (b-d) but for a wrong pose.

5.1.2.2 Segmentation Followed by Recognition

The second approach uses segmentation as a pre-processing step and then analyzes the segment boundaries for matching.

Edge continuity cues in region segmentation: Typically, the continuity constraints are imposed on the segment boundaries - high curvatures are penalized and straight boundaries are promoted. Leung and Malik proposed a pairwise pixel affinity which takes into account intervening gradients between them [49]. N-cuts was used to obtain the final region segmentation. Ren and Malik presented a segmentation scheme in which super-pixels were computed as a pre-processing step for

segmentation [69]. The continuity of super-pixel edges along a segment's boundary were included as part of the segment's goodness value. Yu and Shi generalized the n-cuts algorithm to partition both the pixels and edge elements [96]. The graph nodes corresponding to edge elements are connected by affinities based on continuation. However, obtaining segments that directly correspond to holistic objects is a challenge. Usually, over-segmentation followed by recognition on groups of segments is favored e.g. [56], etc.

Jermyn and Ishikawa proposed an energy function for segmentation which includes both region and boundary cues [40]. The basic idea is to integrate the function along boundaries of segments and choose the segment with lowest energy. There has been related work on integrating segments using region and boundary cues [19, 50].

Part-based Detection: A closely related approach is based on detecting limbs as components shaped as rectangles and combining them using graphs or trees. The rectangles are detected using templates with uniform interior color and contrasting color in the periphery [70, 65, 66]. In [86], the components are combined using a cascade. It is not clear how these techniques could prevent errors due to asymmetric matching - the case shown in Figure 5.1. These methods can easily ignore the extended arm in Figure 5.1(a) and confine themselves to the torso - leading to an erroneous match.

5.1.2.3 Use Edge Continuity during Recognition

The third approach - the one followed here - is to avoid performing segmentation while still taking into consideration edge continuity constraints. Given an image and a pose-contour to be matched, we find the set of gradients in the image that are likely to belong to the subject. If the given pose-contour is correct then this set *must* belong to the foreground. However, this "initial" set might be closely linked with other gradients in the image - which must also belong to the foreground. Edge continuity is used to expand the initial set to include other linked gradients. For the given pose-contour to be a good match to the image, it should match with the expanded set of gradients. The matching is performed using a modified form of chamfer distance. This framework provides a large measure of resistance to spurious matches in the case of highly textured scenes, and to incorrect matches when some poses match only partially with the subject but obtain a high score by avoiding integrating errors in articulated parts of the body (as illustrated in Figure 5.1).

A closely related approach for detecting lakes in satellite imagery was proposed by Elder et.al. [23]. Here, edge continuity constraints are included in a probabilistic model to detect closed contours in edge maps. The authors also describe a method for learning the edge continuity priors in the context of detecting lakes. In our problem, the goal is to match a given set of contours with an image - this is different from the detection problem addressed in [23].

Thayananthan et.al. [82] proposed an improvement to the Shape Context technique by enforcing neighborhood constraints on the matchings between point sets.

They require that neighboring points on the pose-contour be mapped to neighboring points on the image. However, it is not clear whether this would guarantee that the mapped gradients also form a holistic object.

Region-based Segmentation and Recognition: Additionally, there have been many recent studies on linking segmentation and object recognition. Cremers et.al. introduced a variational framework for combining segmentation and recognition [20]. Yu et.al. introduced a generalized version of the normalized-cuts algorithm in which the graph affinities include body-part configuration constraints along with spatial continuity criteria [97]. Borenstein et.al. extended the multiscale segmentation algorithm to enable object recognition by using the segments' saliency as constraints [13, 12]. These approaches employ region-based segmentation and appearance modelling. We complement them by introducing a model for combining edge grouping with contour matching.

5.1.3 Overview of Present Work

Our model for matching a pose-contour to an image combines two measures:

1. The first one measures how well the pose-contour aligns with the gradients in the image. This is computed using an extended form of chamfer matching applied to a continuous gradient magnitude field instead of a discrete edge map. We refer to this as $c_{p \to i}$.

2. The second measures how well the subject's gradients in the image align with the pose-contour. It verifies whether the image gradients underlying the test

pose-contour form a holistic object, or are part of a larger object. This measure is computed from the expanded set of gradients obtained from edge continuity. It is referred to as $c_{i \to p}$.

We propose an edge-affinity model for grouping edge elements in natural images depending upon whether they could belong to the same object. A pair of edge elements have high affinity if their orientations have good continuity and their neighborhoods have similar color statistics. Given an image and a pose-contour to be matched, an initial set of edge elements matching with the pose-contour is obtained. An iterative process is then used to expand this set to include other edge elements having high affinity with its members. The measure, $c_{i \to p}$, is computed from the degree of mismatch between the estimated outline of the subject and the pose-contour being considered.

The pose contours used in the present study were collected as part of a gesture recognition system. The training database consists of 14 gestures performed by 5 subjects (c.f. Section 5.5). The subjects stand upright and the arms are the principal modes of gesticulation. The proposed pose-matching system is tested both with still images and in a gesture recognition application.

We first review work on edge continuity and then describe the edge affinity model. Section 5.3 describes the algorithm for using the edge affinities to compute $c_{i \to p}$. The extended form of Chamfer matching is described in section 5.4.

5.2 Edge Affinity

Two edge elements in a given image are said to have high affinity if they are likely to be part of an object's boundary. This depends upon:

1. The "goodness" of the contour that could pass between them, with the contour's orientation constrained by the orientation of the edge elements.

2. The color statistics in their neighborhoods.

The proposed edge affinity model is presented in stages. First the dependence on the curvature of the contour connecting the two edge elements is described (c.f. eq. (5.2)). Next, the orientation of the edge elements w.r.t. this contour is included (c.f. eq. (5.3)). Finally, color statistics in the neighborhood of the edge elements are factored in (c.f. eq. (5.4)).

5.2.1 Edge Continuity

Given two edge elements, edge continuity criteria measure how likely it is that they are connected. This has been extensively studied in computer vision for detecting salient figures in images and for forming subjective contours [42]. Sha' Ashua and Ullman computed the saliency of edges by building a network of edge elements and use curvature and curvature variations to formulate a measure of saliency [74]. Parent and Zucker used the concept of an osculating circle for edge continuity [58]. Guy and Medioni combined this with tensor voting to obtain saliency maps for the edges in an image [34]. Williams et.al. proposed a stochastic completion

model to compute the probability that a contour connecting one point to another would pass through a given intermediate point. The obtained probability fields show interesting resemblance to subjective contours [90]. Although edge-continuity has been studied in the context of perceptual grouping, we are not aware of any work in linking it with recognition. We use the model proposed by Parent and Zucker for our edge affinity model.

5.2.1.1 Osculating Circles

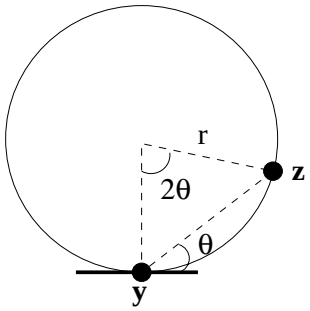


Figure 5.2: Osculating circle given two points **y** and **z** lying on it and the tangent to the curve at **y**.

Edge-continuity constraints typically assume that curves with low curvature are more likely to occur. In the case of [58], given two points and the orientation of the contour at one of them, the most likely contour to pass through them is assumed to be a circle. The reasoning being, for closed contours with fixed lengths, a circle will have minimum curvature. The circle so defined is called an osculating circle. This is illustrated in Figure 5.2 - **y** and **z** are the two given points on the image plane and the orientation of the contour at **y** is fixed. It can be shown that the

radius of the circle - denoted by $r(\mathbf{y}, \mathbf{z})$ - is given by:

$$r(\mathbf{y}, \mathbf{z}) = \frac{\|\mathbf{y} - \mathbf{z}\|}{2 \sin \theta} \tag{5.1}$$

where θ is as shown in Figure 5.2(a). The curvature of a circle is the reciprocal of its radius. The smaller the curvature, the better connected are the two edge elements at $\mathbf{y}$ and $\mathbf{z}$.

Let $e_{\mathbf{y}}$ denote the edge element at $\mathbf{y}$ on the image plane. We denote the affinity between two edge elements $e_{\mathbf{y}}$ and $e_{\mathbf{z}}$ by $a(e_{\mathbf{y}}, e_{\mathbf{z}})$. It's variation w.r.t. $r(\mathbf{y}, \mathbf{z})$ would depend upon the statistics of the curvature of the contours of humans. The computed statistics are local in nature and depend upon the curves typically observed on outlines of cloths. We analyzed the pose-contours of 5 human subjects while performing the "Turn Left" gesture (Figure 5.11). See Appendix A for details. Based on this analysis, the affinity $a(.)$ is formulated as a sigmoidal function of $r(\mathbf{y}, \mathbf{z})$

$$a(e_{\mathbf{y}}, e_{\mathbf{z}}) = \frac{1}{1 + \exp(-\frac{r(\mathbf{y},\mathbf{z})-6}{.9})} \tag{5.2}$$

Figure 5.3 shows a plot of its variation w.r.t. the radius. The subjects' heights in the images in our application varied from 170 to 200 pixel units - the parameters of the function were kept constant for all experiments. For applications with pose-contours of a substantially different scale: (a) The $r(\mathbf{y}, \mathbf{z})$'s can be scaled linearly w.r.t. the subjects' scales, or (b) statistics of the new contour set can be collected and the constraints on $a(.)$ adjusted according to them.

Until now we have ignored the orientation of $e_{\mathbf{z}}$ when computing $a(e_{\mathbf{y}}, e_{\mathbf{z}})$. Let $\hat{n}(\mathbf{z})$ denote the unit vector tangent to the osculating circle at $\mathbf{z}$. If $e_{\mathbf{z}}$ is orthogonal to $\hat{n}(\mathbf{z})$ then the affinity should be 0. On the other hand, the affinity should be

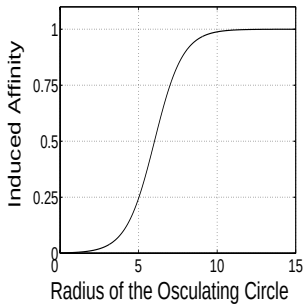


Figure 5.3: Variation of the induced affinity for different radii of the osculating circles.

maximal when $e_{\mathbf{z}}$ is tangential to the osculating circle at $\mathbf{z}$. In general, the affinity is proportional to the magnitude of the normalized projection of $e_{\mathbf{z}}$ onto $\hat{n}(\mathbf{z})$. Let $\Delta I(\mathbf{z})$ denote the gradient vector at $\mathbf{z}$ - it would be perpendicular to $e_{\mathbf{z}}$. Including the orientation factor into the affinity yields

$$a(e_{\mathbf{y}}, e_{\mathbf{z}}) = \frac{1}{1 + \exp(-\frac{r(\mathbf{y},\mathbf{z})-6}{.9})} \left\| \hat{n}(\mathbf{z}) \times \frac{\Delta I(\mathbf{z})}{\|\Delta I(\mathbf{z})\|} \right\| \tag{5.3}$$

5.2.2 Including Color Statistics

The large amount of edge clutter present in natural images makes edge continuity alone unreliable for determining edge affinities. Color statistics in the neighborhoods of the edges form an important low level cue for grouping. In the case of edges bordering an object, only one side of the edge (the foreground side) should have similar colors. The other side, belonging to the background, can have arbitrary colors. Therefore, we collect statistics on both sides of the edges but constrain only the side indicated by a candidate contour to be the foreground. The color statistics are collected by averaging the color in 5×5 windows on either side of the edge elements - see Figure 5.4(a).

We label the sides adjacent to an edge as +ive and −ive depending upon the orientation of ΔI at that point. Accordingly, the color statistics at an edge element $e_{\mathbf{y}}$ are denoted by $\mathbf{c}_+(e_{\mathbf{y}})$ and $\mathbf{c}_-(e_{\mathbf{y}})$. Now consider two edge elements $e_{\mathbf{y}}$ and $e_{\mathbf{z}}$, and, without loss of generality, suppose that the +ive side of $e_{\mathbf{y}}$ belongs to the foreground. When extending the contour from $\mathbf{y}$ to $\mathbf{z}$, the side of the osculating circle corresponding to the +ive side of $e_{\mathbf{y}}$ will be the foreground and hence should exhibit color constancy - see Figure 5.4(b). Depending upon the angle made by $\Delta I(\mathbf{z})$ with the tangent to the osculating circle ($\hat{n}(\mathbf{z})$), one of $\mathbf{c}_+(e_{\mathbf{z}})$ and $\mathbf{c}_-(e_{\mathbf{z}})$ is chosen for comparison with $\mathbf{c}_+(e_{\mathbf{y}})$; in the shown example $\mathbf{c}_-(e_{\mathbf{z}})$ would be chosen. For computing the orientation of $\Delta I(\mathbf{z})$ w.r.t. $\hat{n}(\mathbf{z})$, we compute the cross-product $\Delta I(\mathbf{z}) \times \hat{n}(\mathbf{z})$, which is perpendicular to the image plane. Let $\mathbf{c}_s(\mathbf{y})$ be chosen as foreground. There are two cases:

1. $\Delta I(\mathbf{z}) \times \hat{n}(\mathbf{z})$ points upwards: in this case $\mathbf{c}_s(\mathbf{z})$ should be used for comparison.

2. The cross-product points downwards: in this case $\mathbf{c}_{-s}(\mathbf{z})$ should be used for comparison.

In other words, if $\mathbf{c}_s(\mathbf{y})$ is chosen as foreground then $\mathbf{c}_{s'}(\mathbf{z})$ is chosen for comparison, with $s' = s\ \mathrm{sgn}((\Delta I(\mathbf{z}) \times \hat{n(\mathbf{z})}) \cdot (\hat{k}))$. Here $\hat{k}$ is a unit vector perpendicular to the image plane, pointing upwards.

Let the color statistics chosen at $\mathbf{y}$ and $\mathbf{z}$ be denoted by $\mathbf{c}_{\mathbf{y}}$ and $\mathbf{c}_{\mathbf{z}}$ respectively. We allow for additive Gaussian noise in the color statistics, and correspondingly

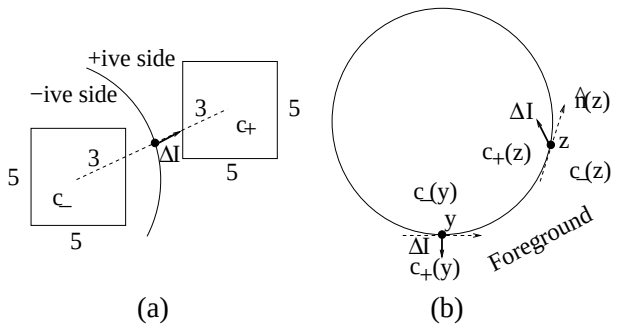


Figure 5.4: (a) Collecting color statistics in 5×5 windows on either side of edge elements. (b) As $c_-(\mathbf{z})$ lies on the foreground side of the osculating circle, it is chosen for comparison with $c_+(\mathbf{y})$.

extend the edge affinity model as:

$$\tilde{a}(e_{\mathbf{y}}, e_{\mathbf{z}}) = \frac{1}{1 + \exp(-\frac{r(\mathbf{y},\mathbf{z})-6}{.9})} \left\| \hat{n}(\mathbf{z}) \times \frac{\Delta I(\mathbf{z})}{\|\Delta I(\mathbf{z})\|} \right\| \exp\left(-\frac{\|\mathbf{c_y} - \mathbf{c_z}\|^2}{\sigma_c^2} \right) \tag{5.4}$$

σ_c was kept at 0.003. The affinity so defined is asymmetric. It is made symmetric by taking the maximum:

$$a(e_{\mathbf{y}}, e_{\mathbf{z}}) = \max\left(\tilde{a}(e_{\mathbf{y}}, e_{\mathbf{z}})\tilde{a}(e_{\mathbf{y}}, e_{\mathbf{z}})\right) \tag{5.5}$$

5.2.3 Using Edge Affinities to Propagate Edges

When a contour is placed on an image, the gradients in the image lying underneath the contour are said to match with it. These are called the *activated* gradients. It is possible that the activated gradients are actually part of a larger object in the image. In this case, they would have high affinities with other gradients not activated by the pose. Let us call these the *propagated* gradients. The activated and propagated gradients together constitute the net saliency induced by the pose on the image. The term *salient* gradients is used to indicate the union of activated and propagated gradients. They would highlight the outline of the object whose edges

were activated by the contour under consideration.

An iterative approach is followed for obtaining the salient gradients, where the previous stage's salient gradients propagate to other gradients through the edge affinities. Let $A^0(e_{\mathbf{y}})$ denote the activation field defined on the image plane - it quantifies the degree of activation of the various edge elements in an image by a contour. This would form the initial saliency field in the iterative process. $A(.)$ ranges over $[0, 1]$. At each iteration, the salient gradients in the neighborhood of an edge element induce saliency to it - the higher the affinity, the greater the saliency induced. For simplicity, we consider only pairwise interactions and use the max operator to combine the saliency induced by the different neighbors of a point. The saliency field at the t^{th} iteration ($t = 0 \ldots \Gamma$) is denoted by $A^t(e_{\mathbf{z}})$, and is computed as

$$A^t(e_{\mathbf{z}}) = E(e_{\mathbf{z}}) \max_{y \in N(\mathbf{z})} \left[a(e_{\mathbf{y}}, e_{\mathbf{z}}) \Psi(A^{t-1}(e_{\mathbf{y}})) \right] \tag{5.6}$$

$E(e_{\mathbf{z}}) \in [0, 1]$ quantifies the confidence of edge element $e_{\mathbf{z}}$ to belong to the foreground. In the absence of additional information, e.g. foreground color statistics, $E(e_{\mathbf{z}})$ is simply the gradient magnitude of $e_{\mathbf{z}}$. $N(.)$ defines an 11×11 neighborhood around a point in the image plane. $\Psi(.)$ is in general a nondecreasing function with range $[0, 1]$. In our implementation it was a step function:

$$\Psi(q) = \begin{cases} 1 & q \geq \delta \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

The threshold helps in reducing computational complexity as points with very low activation can be ignored. In all our experiments, δ was kept constant at 0.005.

For the purposes of illustrating the functionality of the edge affinities, consider the image and its gradient magnitude map shown in Figures 5.5(a) and (b) respectively. We activate a point on the edge of the torso of the subject and extend its edge using the edge affinities. The activated point's coord.s are (128, 105) and it is marked with a circle in Figures 5.5(a) and (b). Thus, the initial saliency field, $A^0(.)$ is as follows:

$$A^0(e_{\mathbf{y}}) = \begin{cases} 1 & \mathbf{y} = (128, 105) \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

Figure 5.5(c) shows the saliency field obtained after 4 iterations, i.e. $A^4(.)$, when the torso side of the initiating edge is made the foreground. Figure 5.5(d) shows the salient points in $A^4(.)$ marked with dots. Figure 5.5(e) shows $A^4(.)$ when the wrong side, i.e. the one on the brick wall, is made the foreground. Figure 5.5(f) shows the salient points in this case. Depending upon the choice of foreground, either the subject or the wall's edges are propagated. Figure 5.6 shows propagation at intermediate stages when the subject's torso is chosen as foreground. Figure 5.7 shows more examples of images and propagations obtained from a single seed edge (marked with a circle) - for these cases the foreground is always chosen to be inside the subject. The edge affinity model is effective in confining the saliency propagations to the subject's edges and prevents the background edges from being highlighted. In the third case, the saliency "jumps" across the subject's sleeve as the edges of the sleeve are parallel and obey constraints on color statistics. Note that we do not expect the whole figure of the subject to be highlighted by just one seed edge. The examples are used to illustrate how the edge affinities characterize the grouping among the

edges.

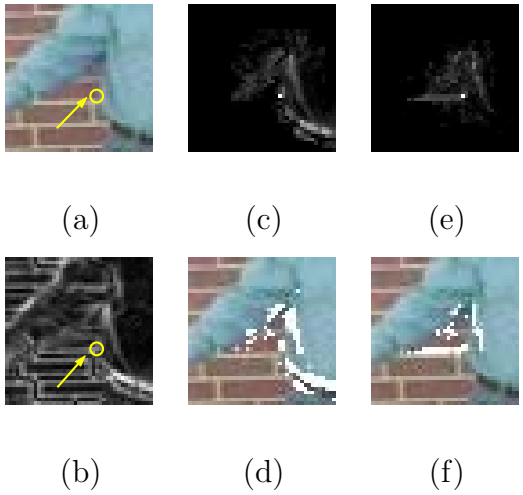


Figure 5.5: (a) Image and (b) its gradient magnitude map with the seed edge element marked with a circle. (c) Saliency field obtained when the side inside the subject is considered foreground - the subject's edges are made salient, (d) to clearly highlight the propagation, points with saliency greater than 0.1 are marked with dots. (e) The case when the side on the brick wall is considered foreground so the wall's gradients are made salient, (f) points with saliency greater than 0.1 marked with dots.

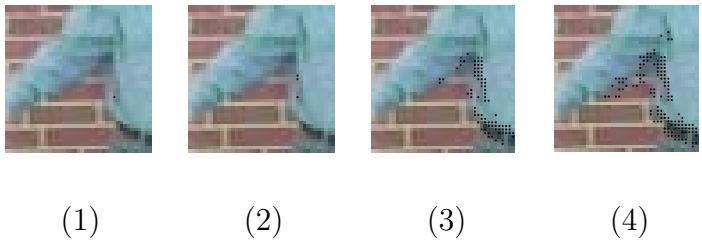


Figure 5.6: Propagation of saliency at different iterations for the image in Fig. 5.5(a) with subject's torso as foreground. Points with saliency greater than 0.1 are marked with dots.

5.3 Computing $c_{i \to p}$

Obtaining the Activation Fields: In most pose tracking and gesture recognition applications, a bootstrap subject-detection phase is used to locate the subject in the field of view. This provides the approximate location and scale of the subject

(1) (2) (3) (4)

Figure 5.7: Images with the initial seed edge element marked with a circle, and the corresponding salient gradients ($A^4(.)$) obtained when the side inside the subject is chosen to be foreground (activated points marked with white dots). Best viewed on color monitor.

for pose matching. However, when a pose-contour is placed on an image, it will not coincide exactly with the subject's gradients in the image. This could be due to variation in subject morphology, apparel, gesticulation style, etc. We allow for Gaussian additive noise in the location of the points on the pose-contours. Each pose p_k is specified as a set of points $\{\mathbf{x}_i^k\}$ outlining the subject's figure in the training data. The Gaussian noise kernel for each $\mathbf{x}_i^k$ follows a multi-variate distribution, with Σ_i^k as the covariance matrix. Let $A_k^0(e_\mathbf{y})$ denote the activation field induced by pose p_k on the image plane. The degree of activation induced by a pose at a point on the image plane is the maximum over the activation induced by individual points of the pose-contour.

$$A_k^0(e_\mathbf{y}) = E(e_\mathbf{y}) \max_{\mathbf{x}_i^k \in p_k} \exp[-(\mathbf{x}_i^k - \mathbf{y})^T (\Sigma_i^k)^{-1} (\mathbf{x}_i^k - \mathbf{y})] \tag{5.9}$$

Figure 5.8 shows examples of activation fields induced by two poses. Here, $E(e_\mathbf{y}) = 1$ for illustrative purposes. The Σ_i^k's are computed from the displacement of different

points on the pose-contours in the training images.

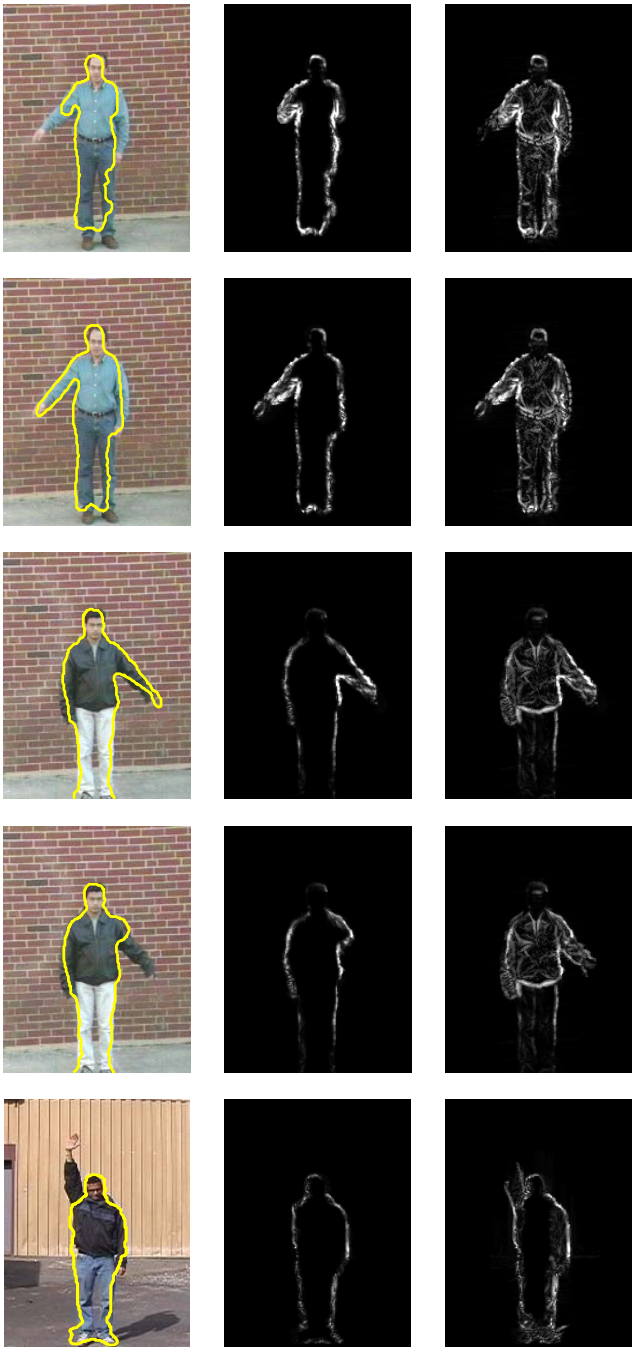
Figure 5.8: Examples of activation fields induced by poses.

The net saliency induced by a pose is obtained by propagating the activation fields in the iterative manner described above. Figure 5.9 shows some examples of images with the pose-contours overlayed, the initial activation fields $A_k^0(.)$'s, and the net saliency fields $A_k^\Gamma(.)$'s. The contours were obtained from the pose database of the gesture recognition system and were manually imposed on the images for illustrative purposes. We see that the objective of highlighting the figure of the subject is achieved. Moreover, in the cases of correct poses, the net saliency fields lie close to the pose-contours whereas the incorrect poses cannot "explain" the net saliency fields.

The next step is to measure the quality of the match between each pose p_k and its net saliency field $A_k^\Gamma(.)$. This is achieved using the Chamfer distance approach. Let $D_k(\mathbf{y})$ denote the distance transform constructed from pose p_k. For the k^{th} pose, $c_{i\to p}^k$ is computed as

$$c_{i\to p}^k = \frac{\sum_{\mathbf{y}} A_k^\Gamma(\mathbf{y}) \exp(-D_k(\mathbf{y}))}{\sum_{\mathbf{y}} A_k^\Gamma(\mathbf{y})} \tag{5.10}$$

$c_{i\to p}$ will be high when the salient gradients in $A_k^\Gamma(.)$ are located close to the pose-contour p_k - this corresponds well with our intuitive notion of a good match.



(a) (b) (c)

Figure 5.9: Examples of saliency fields obtained upon propagation: (a) images with the poses overlayed, (b) initial activation fields ($A_k^0(.)$'s), and (c) net saliency fields -$A_k^\Gamma(.)$'s ($\Gamma = 7$). In case of correct poses, the propagated gradients are close to the original pose-contour whereas the incorrect poses fail to account for all the propagated gradients.

5.4 Extended Chamfer Matching for Computing $c_{p \to i}$

In classical chamfer matching, an image is first reduced to a map of feature points and a distance map is constructed from this feature map. The pose-contour to be matched is placed on the distance map and the distances are integrated along the contour. If the pose-contour matches well with a subset of features in the image then this integral would be small. The feature maps could be edge maps generated by thresholded gradient magnitudes, etc.

The basic form of chamfer matching is limited because:

1. It is difficult to choose a threshold so that only the subject's edges are present in the feature map.

2. The method does not incorporate any prior information about the subject's appearance. In many applications, the subject's color profile does not change during a session. Therefore, color statistics could be used to eliminate some of the background clutter.

3. The integration of the errors (distances) is unweighted - i.e. the method does not take into account any prior knowledge about the uncertainty in the location of different points on the pose-contours. Human arms are the principle modes of gesticulation, causing the contour points on the arms to have the greatest errors in location. However, in spite of being difficult to match, the arms are the key distinguishing features between poses. Therefore, we would like to give less weight to errors at points on the arms.

We address the first two of the issues by using an analog form of the feature map, denoted by $E(e_{\mathbf{y}})$, which quantifies the confidence that the edge element at a point $\mathbf{y}$ on the image plane belongs to the subject. This is computed using the subject's color statistics collected at the person-location phase. In the absence of such information, $E(.)$ is the gradient magnitude field. The issue of weighing the errors in location is handled using Gaussian kernels on the points of the pose-contour. The covariance matrix for the Gaussian kernel at point $\mathbf{x}_i^k$ is Σ_i^k - the same as the one used for computing the activation values.

Let s_i^k denote the confidence value for point $\mathbf{x}_i^k$ on pose-contour p_k to be present in a given image. It is computed as a weighted average of the confidence values for $\mathbf{x}_i^k$ to be located at different points on the image plane. This is simply:

$$s_i^k = \sum_{\mathbf{y}} E(e_{\mathbf{y}}) \frac{e^{(-(\mathbf{x}_i^k - \mathbf{y})^T (\Sigma_i^k)^{-1} (\mathbf{x}_i^k - \mathbf{y}))}}{\sqrt{2\pi}|\Sigma_i^k|} \tag{5.11}$$

We might also take into account the information provided by the orientation of the pose-contour, denoted by $O(\mathbf{x}_i^k)$. A point $\mathbf{x}_i^k$ on a pose-contour can correspond to a point $\mathbf{y}$ on the image plane only if the orientation of gradient at $\mathbf{y}$ $(\Delta I(\mathbf{y}))$ is similar to the orientation of the pose-contour at $\mathbf{x}_i^k$. For this, $s_i^k(R)$ can be expanded to include a function $\phi(\Delta I(\mathbf{y}), O(\mathbf{x}_i^k))$ which quantifies the similarity in orientation.

$$s_i^k = \sum_{\mathbf{y}} E(e_{\mathbf{y}}) \phi\left(\Delta I(\mathbf{y}), O(\mathbf{x}_i^k)\right) \frac{e^{(-(\mathbf{x}_i^k - \mathbf{y})^T (\Sigma_i^k)^{-1} (\mathbf{x}_i^k - \mathbf{y}))}}{\sqrt{2\pi}|\Sigma_i^k|} \tag{5.12}$$

$\phi(.)$ is defined as

$$\phi(\mathbf{v}_1, \mathbf{v}_2) = \left| \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|} \right| \tag{5.13}$$

This gives the confidence value for individual points on the pose-contours. The

$c^k_{p\to i}$'s are obtained by averaging over the pose-contour's points.

$$c^k_{p\to i} = \frac{1}{N_k}\sum_i^{N_k} s^k_i \tag{5.14}$$

where N_k is the number of points on the pose-contour p_k.

Net Confidence for a Pose

The net confidence for a pose is denoted by c_k and is computed as

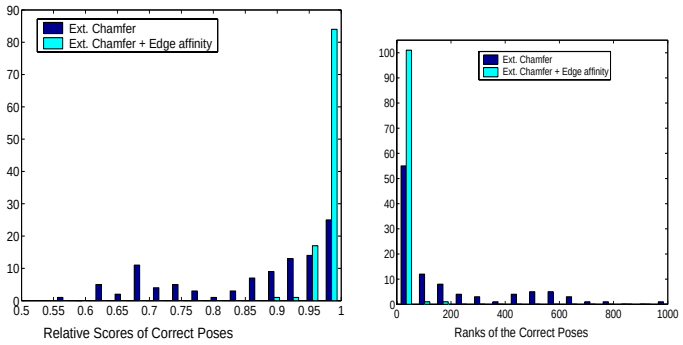
$$c_k = c^k_{p\to i} + c^k_{i\to p} \tag{5.15}$$

5.5 Experimental Results

5.5.1 Still Images

We tested the pose matching model with 103 natural images to observe the improvement due to the edge affinity model and $c^k_{p\to i}$. The test images had cluttered backgrounds, including brick walls, grass, parking lots, etc. The pose-database consisted of 1847 poses performed by 5 subjects (a subset of the database is shown in Figure 5.11). The poses in the database are registered to one another w.r.t. the heads of the subjects. The test images were generated by 4 subjects, 3 of whom were not present in the pose-database; the one common subject was wearing different clothing. The height of the subjects in the images varied from 170 to 200 pixel units.

In pose tracking and gesture recognition applications, the objective of pose-matching is to generate likelihoods for the poses, which are then used by methods like Hidden Markov Models (HMMs), etc. to perform the actual tracking or recognition.



(a) (b)

Figure 5.10: (a) Histogram of the relative confidences of the correct poses - the distribution moves substantially towards 1 upon including edge affinities ($c_{i \to p}$). (b) Histogram of the ranks of the correct poses - the distribution has a significant shift towards 1 upon inclusion of $c_{i \to p}$.

Therefore, the following metrics were used for evaluating the pose matching:

1. The relative confidence of the correct pose - given by the ratio of its confidence value w.r.t. the highest matched (possibly incorrect) pose. This should be as close to 1 as possible, and would ensure that the correct pose is assigned a high confidence.

2. Rank of the correct pose based on its confidence value. This would ensure that the correct pose "stands out" in the pose-database.

For each test image, the confidence values were computed for all poses in the pose-database. The correct pose in the database was manually selected and its ranking and relative confidence were noted. In the case of multiple correct poses, the best match was considered.

Table 5.1 shows the frequency of occurrence of the relative confidences in several high confidence ranges, with and without the inclusion of edge affinities

($c^k_{i\to p}$'s). The frequency counts are boosted by more than 2.5 times when $c^k_{i\to p}$'s are included. With the inclusion of $c^k_{i\to p}$'s, the correct pose had relative confidence greater than 0.95 in all but 3 test images. Moreover, the mean of the relative confidences of the correct poses increased from 0.865 to 0.987 and the standard deviation decreased from 0.127 to 0.018 - an improvement of an order of magnitude.

Range of rel. conf.	$c^k_{p\to i}$	$c^k_{p\to i} + c^k_{i\to p}$
$= 1$	16.5%	44.6%
$[.975, 1]$	23.3%	76.7%
$[.95, 1]$	32.0%	97.1%

Table 5.1: Frequency of occurrence of relative confidences of correct poses in some ranges.

Figure 5.10(a) shows the histogram of the relative confidences of the correct poses for the test images, with and without $c^k_{i\to p}$'s. There is a clear shift in the distribution towards 1 upon inclusion of edge affinities. In all but 6 cases, there was an improvement in the relative confidences upon including edge affinities.

Table 5.2 shows the frequency of occurrence of the ranks of the correct poses in low rank ranges. The frequency counts improve by more than two times upon including $c^k_{i\to p}$'s. Moreover, with 0.94 probability the correct poses are ranked in the top 30 matches as opposed to .40 without $c^k_{i\to p}$. Figure 5.10(b) shows the histogram of the ranks of the correct poses, with and without the inclusion of $c^k_{i\to p}$'s. There is a significant shift in the distribution towards 1 upon inclusion of edge affinities.

Range of rank	$c^k_{p\to i}$	$c^k_{p\to i} + c^k_{i\to p}$
$= 1$	16.5%	44.6%
$[1, 10]$	32.0%	79.6%
$[1, 20]$	38.8%	90.3%
$[1, 30]$	39.8%	94.2%

Table 5.2: Frequency of occurrence of the ranks of correct poses in some ranges.

Thus, the edge affinity model significantly improves the confidences of the correct poses w.r.t. the rest of the pose-database.

5.5.2 Gesture Recognition Results

The contour matching model was used for the gesture recognition application described in [24, 78]. We considered 11 of the 14 gestures in the database as the other 3 required motion features for good discrimination. For each gesture we collected 25 sequences, 5 of which were used as exemplars, and 20 for testing. The classification accuracy was 68.64% when only $c^k_{p\to i}$'s were used. This improved to 79.55% when edge affinities ($c^k_{i\to p}$'s) were also included[1]. The confusion matrix for the recognition of the test sequences, with and without $c^k_{i\to p}$'s, is given in Table 5.3. Inclusion of edge affinities improves the recognition rates of the gestures.

[1]In practice we would include motion features to improve the recognition accuracy; the results presented here are based only on shape.

	Turn-Left	Turn-Right	Flap	Stop-Left	Stop-Right	Stop-Both	Attention-Left	Attention-Right	Attention-Both	Start Engines	Speed Up
Turn-Left	**13/16**	0/0	3/3	0/0	0/0	1/0	0/0	0/0	1/0	1/1	1/0
Turn-Right	0/0	**3/8**	16/12	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0
Flap	0/0	0/0	**20/20**	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Stop-Left	1/0	0/0	0/0	**9/13**	0/1	10/5	0/0	0/0	0/0	0/0	0/1
Stop-Right	0/0	0/0	0/0	0/0	**14/17**	6/3	0/0	0/0	0/0	0/0	0/0
Stop-Both	0/0	0/0	0/0	0/0	0/0	**20/20**	0/0	0/0	0/0	0/0	0/0
Attention-Left	1/0	0/0	0/0	0/0	0/0	0/0	**18/18**	0/0	1/1	0/1	0/0
Attention-Right	0/0	0/0	0/0	0/0	0/0	0/1	0/0	**18/18**	2/1	0/0	0/0
Attention-Both	0/0	0/0	0/0	0/0	0/0	0/0	1/0	6/4	**13/16**	0/0	0/0
Start Engines	0/0	0/0	0/0	3/0	0/0	4/2	0/0	0/0	1/0	**12/18**	0/0
Speed Up	1/1	0/0	0/0	0/0	0/0	2/3	3/1	0/0	3/4	0/0	**11/11**

Table 5.3: Confusion matrix for the gesture recognition. Entry a/b in the i^{th} row and j^{th} column indicates that a sequences actually depicting gesture i got classified as gesture j when only $c^k_{p\to i}$'s were used, and b indicates the number of classifications when edge affinities were also considered. Correct classifications are indicated in bold face.

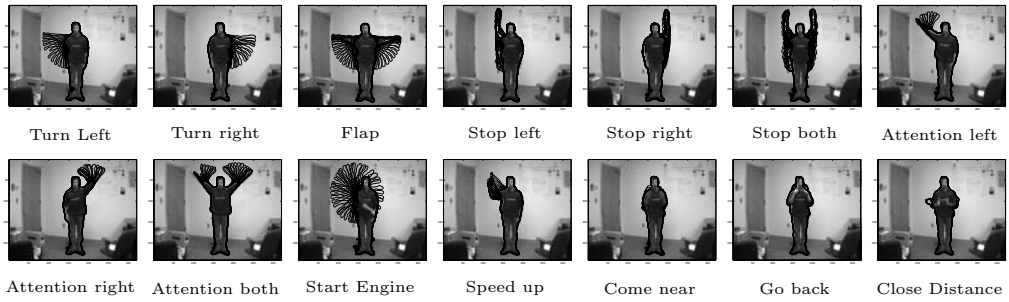


Figure 5.11: Shape exemplars for each gesture overlayed over the images

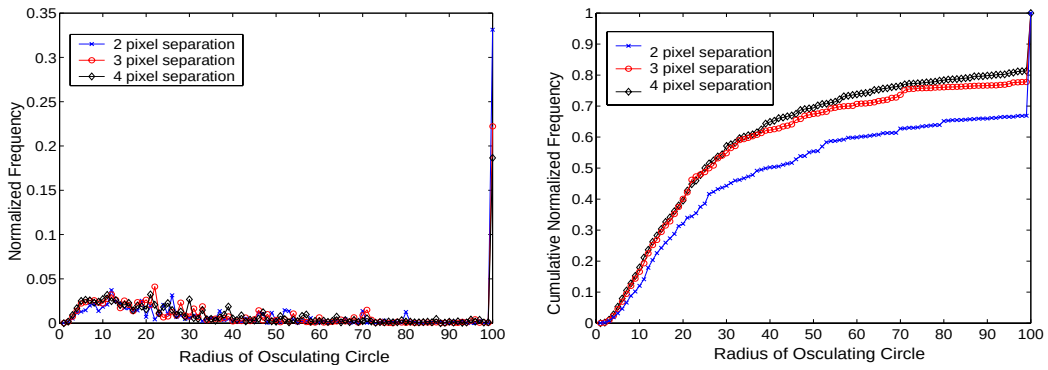
5.6 Summary

We presented a model for combining edge-continuity with contour matching, and illustrated its utility in the context of human pose matching. The experiments indicate that the model is able to characterize the inherent grouping of the edges - e.g. Figures 5.5 and 5.7. The tests show that the use of edge affinities leads to significant improvements in matching. This demonstrates the importance of perceptual organization for object recognition.

Appendix A: Dependence of Edge Affinity on Radius of Osculating Circle

The edge affinity function, $a(\mathbf{y}, \mathbf{z})$'s, variation w.r.t. $r(\mathbf{y}, \mathbf{z})$ would depend upon the statistics of the curvature of the contours of humans. These statistics are local in nature and depend upon the curves typically observed on outlines of cloths. We analyzed the pose-contours of 5 human subjects while performing the "Turn Left" gesture. For this, the radii of the osculating circles connecting pairs of points

along the pose-contour were computed. The distance between the points in each pair, i.e. $\|\mathbf{y} - \mathbf{z}\|$, was kept at 2, 3 and 4 pixel units. Figure 5.12(a) shows the normalized frequency of occurrence of osculating circles of different radii for each separation distance. The radii values are in pixel units and were capped at 100 units. There are two modes in the distribution, the first one is formed by radii between 5 and 20 units, and the second mode corresponds to straight segments with radii greater than 100 units. Figure 5.12(b) shows the cumulative normalized frequency of the same values.



(a) (b)

Figure 5.12: (a) The normalized frequency of occurrence of osculating circles of different radii, for pairs of points along contours of whole body. The three plots correspond to point-pairs separated by 2, 3 and 4 pixel units. (b) The cumulative normalized frequency of occurrence of osculating circles of different radii, for pairs of points along contours of whole body. The three plots correspond to point-pairs separated by 2, 3 and 4 pixel units.

The following observations can be made regarding the dependence of $a(e_{\mathbf{y}}, e_{\mathbf{z}})$ on $r(\mathbf{y}, \mathbf{z})$:

- $a(.) \in [0, 1]$.
- As curvature increases, $a(.)$ rapidly tends to 0. $a(e_{\mathbf{y}}, e_{\mathbf{z}}) \approx 0$ for $r(\mathbf{y}, \mathbf{z}) \leq 3$.

- $a(e_{\mathbf{y}}, e_{\mathbf{z}}) \rightarrow 1$ as $r(\mathbf{y}, \mathbf{z}) \rightarrow \infty$. As curvature becomes 0, $a(.)$ asymptotically approaches 1.

- Nearly 90% of the observed radii were ≥ 10. To ensure that a majority of the edge elements on the subjects' outlines are strongly linked, we kept $a(e_{\mathbf{y}}, e_{\mathbf{z}}) \approx 1$ for $r(\mathbf{y}, \mathbf{z}) \geq 10$.

- The edge affinity is computed in an 11×11 neighborhood around each pixel. Therefore, the maximum value of $\|\mathbf{y} - \mathbf{z}\|$ is $5.5\sqrt{2}$. To allow for some joint articulation, $a(e_{\mathbf{y}}, e_{\mathbf{z}})$ was fixed at 0.5 for 90° bends, i.e. $\theta = 45°$. For $\|\mathbf{y} - \mathbf{z}\| = 5.5\sqrt{2}$, this would correspond to an osculating circle with radius ≈ 6 pixel units. To allow for such bends $a(e_{\mathbf{y}}, e_{\mathbf{z}}) = 0.5$ for $r(\mathbf{y}, \mathbf{z}) = 6$.

The affinity $a(.)$ is formulated as a sigmoidal function of $r(\mathbf{y}, \mathbf{z})$ - the values of the parameters are determined from the mentioned constraints.

$$a(e_{\mathbf{y}}, e_{\mathbf{z}}) = \frac{1}{1 + \exp(-\frac{r(\mathbf{y},\mathbf{z})-6}{.9})} \tag{5.16}$$

Figure 5.3 shows a plot of this function.

Chapter 6

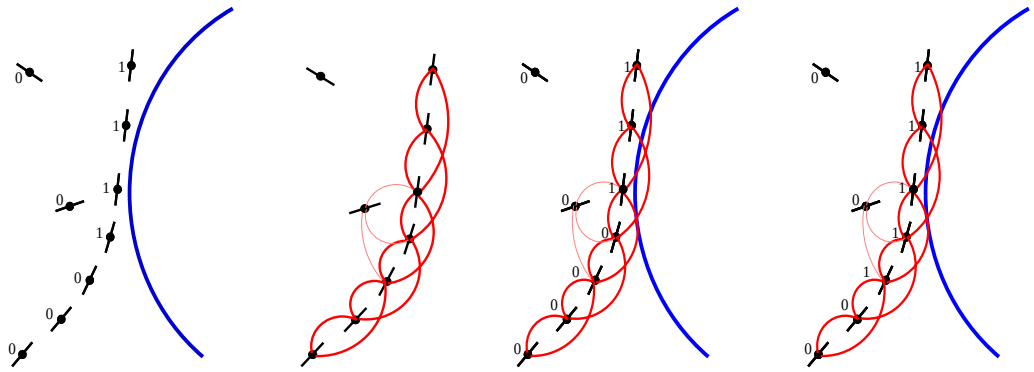
Edge Continuity for Human Detection

Chapter 5 introduced an edge continuity model and employed it for pose-matching. This model is extended by coupling edge continuity and contour matching in a feedback loop, formulated as an energy optimization problem. The approach is illustrated with a human detection application.

Consider the set of edge elements located on the image plane shown on in Figure 6.1(a), and a probe contour matching with them. The edge elements are labelled, '1'-matching and '0'-not matching, based on their proximity and orientation relative to the contour. However, the edges have mutual affinities based on the image structure - Figure 6.1(b). The affinities constrain the labelling. Pairs of edges having high affinities must be assigned similar labels. The contour has the option of either matching with a smaller set of edges at a cost - Figure 6.1(c), or violating some of the edge affinities - Figure 6.1(d). The tradeoff between these two options forms the basis for the feedback loop, formulated as energy optimization.

6.1 Markov Random Field on Edge Elements

Consider a set of edge elements $\{e_i\}_{i=1}^{N}$ on the image plane. A probe contour, C, placed on the image plane induces saliency on the edges. This may be considered as assigning label $l_i \in \{0, 1\}$ to edge e_i, where $l_i = 1$ if the edge is made salient and



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

Figure 6.1: Illustration of edge continuity and contour matching. (a) Edge elements and a probe contour. (b) Edge affinities - strong affinities shown with thick lines. Constrained by the affinities, the contour can either (c) match with a smaller set of edges, or (d) violate some of the affinities by assigning unequal labels.

$l_i = 0$ otherwise. The labels l_i are random variables defined on the edges, forming a field $\mathcal{L}$. The goodness of a contour's match with the edges is determined by the joint likelihood function of the labels $p(l_1, \ldots, l_N)$. This consists of two types of factors:

1. **Single Variable:** A likelihood function, $p_i(l_i)$, is defined for each edge that determines how likely e_i is to be assigned label l_i. Edges located close to the contour and oriented parallel to its local tangent have higher likelihood of being made salient, i.e. $p_i(l_i = 1)$ is high.

2. **Pairwise:** For each pair of neighboring edge elements, e_i and e_j, a likelihood function $p_{ij}(l_i, l_j)$ determines the joint likelihood of their labels. If e_i and e_j have high affinity, $p_{ij}(.)$ constrains them to be assigned similar labels - either both should be made salient or none.

The likelihood of a label l_i of an edge e_i given the labels of the rest of the edges is denoted as $p(l_i|\mathcal{L} - l_i)$, where $\mathcal{L}$ is the entire label field $l_1, \ldots, l_N$. It can be

shown that $p(l_i|\mathcal{L} - l_i)$ is solely determined by e_i's neighbors, $\mathcal{N}_i$

$$p(l_i|\mathcal{L} - l_i) = p_i(l_i|\mathcal{N}_i) \tag{6.1}$$

This forms the basis for the Markovian property of the field of labels. Due to Markov-Gibbs equivalence [32], the joint likelihood of the labels can be modelled as

$$p(l_1, \ldots, l_N) = \frac{1}{Z} \exp\left(-E(l_1, \ldots, l_N)\right) \tag{6.2}$$

where $E(l_1, \ldots, l_N)$ is a Gibbsian energy function defined on the labels and Z is the partition function.

$$E(l_1, \ldots, l_N) = \sum_{i=1}^{N} E_i(l_i) + \Phi_{\text{EA}} \sum_{i<j} E_{ij}(l_i, l_j) \tag{6.3}$$

The single variable terms, $E_i(l_i)$, depend upon the proximity and orientation of the edges relative to the contour. The pairwise terms, $E_{ij}(l_i, l_j)$, depend upon the affinities between the edges. The parameter Φ_{EA} determines the relative importance of single and pairwise energies.

Finding the labelling with maximum likelihood is equivalent to minimizing the energy function $E(.)$ defined on the labels. Markov-Gibbs equivalence and energy minimization have been extensively employed in computer vision for image segmentation, texture analysis, denoising, etc. See [32] for a tutorial on MRFs. Next, we describe the definition of single variable and pairwise energy terms.

6.1.1 Single Variable Terms

For an edge e_i, $E_i(l_i = 0)$ is high if e_i is located close to the contour and oriented parallel to its local tangent; this favors e_i to be salient. On the other hand,

if e_i is located far off from the contour then $E_i(l_i = 1)$ is high, favoring e_i to be not salient. Let δ_i be the distance of edge e_i from the nearest point on the contour, $\|\Delta I(e_i)\|$ the gradient magnitude at edge e_i, and let ϕ_i be the angle of e_i w.r.t. the nearest point on the contour.

$$E^i(l_i) = \begin{cases} \max(\Phi_{\mathrm{d}} - \delta_i, 0)\|\Delta I(e_i)\| \cos\phi_i & : \quad l_i = 0 \\ \max(\delta_i - \Phi_{\mathrm{d}}, 0)\|\Delta I(e_i)\| & : \quad l_i = 1 \end{cases} \tag{6.4}$$

where the parameter Φ_{d} determines the extent of the "spatial spread" of a contour's match with image edges.

6.1.2 Pairwise Terms

Let $a(e_i, e_j)$ be the affinity between edges e_i and e_j. The two variable terms are defined as

$$E^{i,j}(l_i, l_j) = \begin{cases} a(e_i, e_j) & : \quad l_i \neq l_j \\ 0 & : \quad l_i = l_j \end{cases} \tag{6.5}$$

Higher the affinity, more the energy assigned to dissimilar labelling of neighbors.

6.1.3 Energy Minimization

MRF energy minimization has been studied in computer vision for various applications [81], including image registration [10], texture modeling [33], image labelling [18], interactive photo segmentation [72], model-based image segmentation [46]. The most popular and successful approaches include: Graph Cut [15, 45, 14], Loopy Belief Propagation (LBP) [26], and Tree-Reweighted Message passing (TRW) [88]. In a comparative study of energy minimization algorithms for low-

level vision tasks, the Graph Cut algorithm achieves some of the best results and is very efficient [81, 44]. In particular, Graph Cut is guaranteed to compute a globally optimum solution for binary labelling problems with *regular* energy functions [14]

$$E_{ij}(\alpha, \beta) + E_{ij}(\beta, \alpha) \geq E_{ij}(\beta, \beta) + E_{ij}(\alpha, \alpha) \tag{6.6}$$

where α and β are two labels. There are no constrains on the single variable terms. It is easy to see that the pairwise function defined in eq. (6.5) is regular. Therefore, we employ Graph Cut for the optimization.

For a given probe contour, the optimum value of the energy, E^*, would correspond to the goodness of the best possible labelling of the image edges. A high value of E^* indicates that the probe contour is matching with only a subset of the edges of an object in the image. This would imply that the detection corresponding to the match has low likelihood. Next, we describe the application of edge continuity MRF for human detection.

6.2 Human Detection

A large number of approaches have been proposed for human detection in images, see [31] for a survey. We identify three broad categories:

1. Many methods create a database of whole-body contours during training and match these with image edges for detection, e.g. [25, 31, 84], etc. A number of approaches use an articulated model to generate outlines of the object to be detected and then match these with image edges, e.g. [98], etc.

2. A number of studies detect whole-body figures by characterizing edges within sub-windows of a given image and analyzing the obtained features. The edges are represented using Haar-like features e.g. [87], etc., histograms of oriented gradients e.g. [22, 99], etc. Experimental results indicate that these methods are both efficient and effective.
3. Some studies advocate a bottom-up part-based approach to address occlusions and to reduce computational complexity, e.g. edge features are used to characterize parts of the human figure in [54, 52, 71, 92, 93], etc. A closely related group of approaches use region features, e.g. [66, 70], etc.

Scene-geometry has been used to aid object-recognition in [83, 38], etc. It is shown to be useful for eliminating false alarms having scales and/or locations that are incongruous with the scene. This study focuses on using edge grouping in natural images as a constraint on detection.

We employ the Histograms of Gradients (HoG) algorithm [99] for computing an initial set of detections, which are analyzed using edge continuity MRF. HoG is very efficient, enabling a dense scan of the images for instances of humans. It has been shown to be effective in human detections, with very low false detection rates. For instance, in our experiments, each image was densely scanned to produce nearly 48000 overlapping windows of varying scales. The false detection rate of the HoG algorithm was very low - an average of 13 false candidates were observed per image, giving a rate of $\frac{13}{48000} \approx 3 \times 10^{-4}$. As will be shown in the experiments, edge affinities further reduced this false detection rate by nearly 50% while still maintaining the correct detections.

6.2.1 Histograms of Gradients Detector

The HoG detector takes an image window as input and estimates whether the window could contain an instance of a human. The features consist of histograms of gradients computed within patches inside the window. The original detector proposed by Dalal and Triggs [22] employed Support Vector Machines (SVMs) for the classification. Zhu et al. extended this by using a boosting-based classifier that increased the efficiency while maintaining performance [99]. In our experiments, the HoG detector's parameters were set so as to ensure very low false rejects. Please see Appendix A for details of the implementation.

6.2.2 Analysis of HoG Detections

Each detection computed using the HoG algorithm is analyzed with edge continuity. We employed the hierarchical contour matching approach proposed in [31] to compute the most likely human contour for each HoG detection. The contours employed to build the hierarchy were obtained from the MIT pedestrian database. In [31], the goodness of a contour's match is measured using Chamfer distance. The original approach is capable of efficiently searching for humans across scales. For efficiency, we restricted the scale of the search using the size of the detection window computed by HoG. Let C^* be the best matching human contour computed, with s_{Cham} the Chamfer match score. An edge continuity - MRF is constructed for C^*, with the edge elements located in the neighborhood of the detection window. Let E^* be the optimal energy obtained after minimization. The final score for the detection

window is defined as a simple linear summation of the scores

$$s_{\text{Cham}} + \exp(-E^*)$$

The higher the score, greater the estimated likelihood of the detection.

6.3 Experiments

The Edge Affinities for Contour Matching (EACM) was tested with a set of images containing humans in outdoor environments. The data-set consisted of 28 images recorded by a camera mounted on a mobile robot navigating in a wooded scene, and 25 images downloaded from the Internet. There were a total of 64 instances of humans in the images. The images had substantial edge clutter due to the presence of trees, shrubs, etc. The subjects' figures in the images were of varying scales.

The HoG detector was trained on the INRIA data-set [22] - the details are described in Appendix A. It was used to scan each image with overlapping windows of varying scales. The total number of image-windows scanned for each image was nearly 48000. The HoG detector detected 60 of the human instances present in the test data-set. In spite of the dense scanning of the images and the presence of edge-clutter, the detector produced only 441 false alarms - corresponding to a false-alarm rate of 3×10^{-4} per image-window. Upon analyzing the candidate detections using EACM, 58 human instances were correctly detected with a reduction to 209 false alarms. Thus, the number of false alarms was reduced by $1 - \frac{209}{441} \approx 50\%$ while eliminating only two of the correct detections obtained using HoG. Figure 6.2

shows the ROC plots for EACM and the case when only Chamfer-distance score, C, is considered. It indicates that EACM significantly reduces the number of false alarms. Moreover, the ROC plot for Chamfer-distance score indicates that Chamfer-distance alone is unable to improve on the results of HoG. Figure 6.3 shows some test images with the candidate detections obtained using the HoG algorithm and the result of postprocessing with EACM.

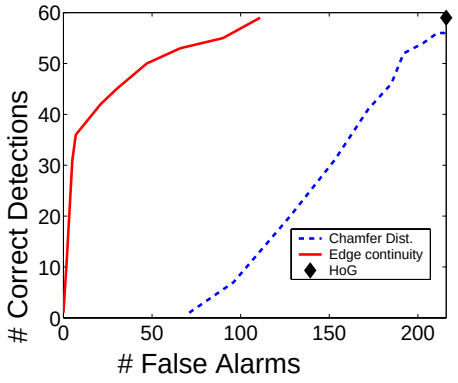
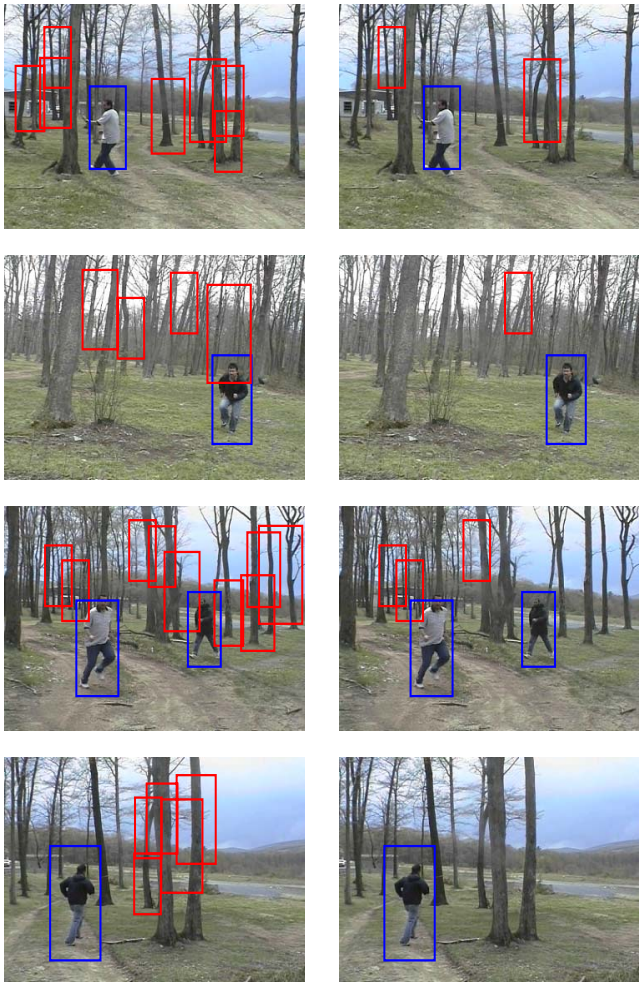


Figure 6.2: ROC plots for EACM - edge affinity coupled with contour matching (in solid-red), and Chamfer-distance score alone (in dashed-blue).

We also created a data-set of 51 color images from the CAVIAR video database which is recorded in an indoor environment [3]. Cases in which the subjects' heights were less than 30 pixels were ignored - the image gradients obtained for these cases were not distinct enough for applying the edge affinity model. The images had a total of 165 instances of humans. HoG detected all of the human instances, with 239 false detections. After post-processing with EACM 154 of the humans were detected($\frac{154}{165} \approx 93\%$), with 168 false alarms (reduction of $1 - \frac{168}{239} \approx 30\%$). Figure 6.4 shows the ROC plots for EACM and the case when only Chamfer-distance score is considered. The plots indicate that EACM again reduces false alarms.



(a) (b)

Figure 6.3: (a) Image with candidate detections produced by the HoG algorithm, (b) detections obtained after post-processing with EACM. Correct detections are marked in blue and false detections in red. (Best viewed in color.)

Figure 6.5 shows some test images with the candidate detections obtained using the

HoG algorithm and the result of considering edge affinities.

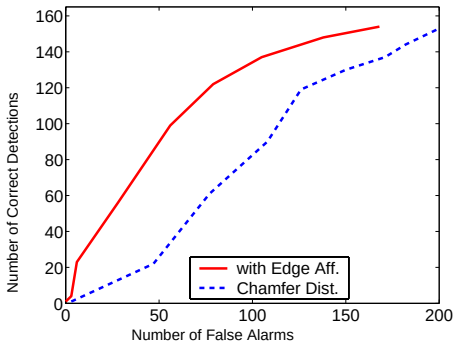
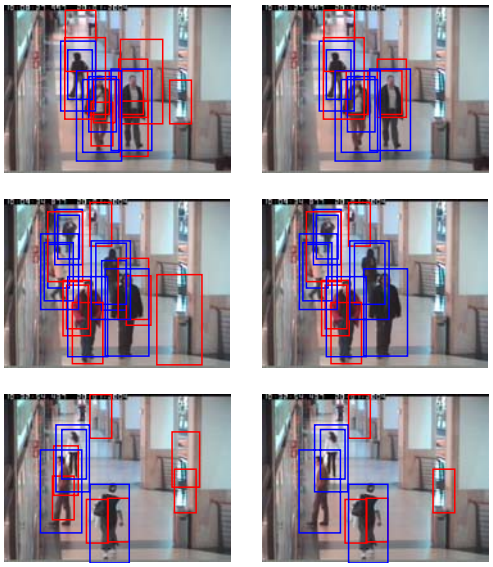


Figure 6.4: ROC plots for edge affinity coupled with matching (in solid-red), and Chamfer-distance score alone (in dashed-blue).



(a) (b)

Figure 6.5: (a) Image with candidate detections produced by the HoG algorithm, (b) detections obtained after combining edge affinities with matching. Correct detections are marked in blue and false detections in red. (Best viewed in color.)

Chapter 7

Summary and Potential Research Directions

We presented a Bayesian model for recognizing ballistic movements such as reaches and strikes based on insights provided by psycho-kinesiological studies. Explicit consideration of the ballistic dynamics enables generalization over target locations and directions of movement. The test results indicate that the approach is robust to changes in camera viewpoint, stylistic variations and subject's morphology and pose w.r.t. the camera.

The second contribution was a model for combining edge-continuity with contour matching. Its utility was illustrated in the context of human pose matching in gesture recognition and human detection. The results indicate that edge affinities result in significantly improved performance.

Next, we describe some potential directions of research.

7.1 Ballistic Movement Model

7.1.1 Temporal Segmentation

The criterion for temporal segmentation of continuous video sequences is that the dynamics within each movement segment must be constant. In the implementation tested in the experiments, the dynamics was represented by the direction of movement, and the segmentation criterion was that the direction of motion within a

movement segment must be consistent. The formulation in eq.(3.8) is quite general and it is possible to extend this further.

For example, a strike movement segment may be erroneously merged with a succeeding reach movement if its duration were very small and the direction of movement similar to that of the reach movement. Such errors may be avoided if the segmentation criterion were to constrain that a segment should not have two markedly different acceleration phases within it. Suppose we were to define the dynamics as $B_i = \langle \theta_i, s_i \rangle$, where $s_i = 1$ if the speed is high - akin to a strike movement, and $s_i = 0$ if the speed is low - akin to reach movements. Let $h(t)$ be the estimate computed by the reach/strike classifier, $h(t) = 1$ for strike movements and $h(t) = 0$ for reach movements. The i^{th} segment's goodness may be defined to depend upon

$$\sum_{t=t_s^i}^{t_e^i} \lfloor s_i = h(t) \rfloor$$

7.1.2 Action Recognition

An action may be considered as a sequence of movements. For example, the action "pick up the book" would consist of a reach-to-grasp movement of the hand to the book, some small movements during grasping, and another movement to move the book up. It is possible to employ Markov Models (possibly with hidden states) to recognize such actions, where each state would correspond to one movement. This is very similar to the SLDSs [68]. While doing so, it is possible to include global parameters such as speed of the action, and the target of the movements, e.g., the

book's location.

7.1.3 Styles of Actions

A number of psychological studies have reported that variations in dynamics are governed to a large extent by variations in the objective of the movement [51, 80]. For instance, when picking up objects, the arm's joint angles at the end of the reach-to-grasp depend upon parameters such as the object's weight and fragility. State-of-the-art image processing techniques lack the accuracy for such subtle measurements [53], but it is possible to explore this concept with motion capture data. This would have applications in

- medical diagnostics: movement styles are symptomatic of the early onset of certain diseases [79].
- automated or assisted coaching for sports: detailed analysis of movement patterns may assist in improving the efficiency of movements.

Another potential application is to employ video-based movement analysis along with a marker-based motion capture system. This may reduce the cost of the motion capture system by lowering the required frame rate and signal to noise ratio in localization.

The movement style may be observed through:

- The pose of the person towards the end of the movement. This includes gait as well as the arm's joint angles.

- The dynamics, in terms of B_i. Local Taylor coefficients of the MJM polynomials at different points along the trajectory would indicate subtle parameters such as acceleration and deceleration.

- The inter-joint coordination during the movement. The onset and end of rotations of different joints depends upon the style of the movement [60]. This is especially evident when the movement speeds and forces are varied.

7.1.4 Generating Animations

It is possible to employ the dynamical model for the inverse problem of generating animations of ballistic movements. State-of-the-art generative approaches use a data-driven paradigm, including [68]. Novel movements are generated by interpolating and extrapolating from specified examples. A better dynamical model would enable more realistic interpolations between example movements. If the trajectory of the two hands and the head were given through examples, then novel whole-body trajectories may be generated by the following general steps:

1. Vary the dynamical parameters of the trajectories to get the hand and head trajectories for the target movement to be generated.

2. A number of movement analysis approaches employ manifold techniques to compute correlations between the trajectories of various body-parts [53]. Employ the manifold and inverse kinematics to compute the whole-body poses from the hand and head trajectories.

The parameters of the dynamical models such MJM, MTCM, would enable systematic variation of the dynamics, which in turn would control the traversal on the manifold of poses.

7.2 Edge Continuity

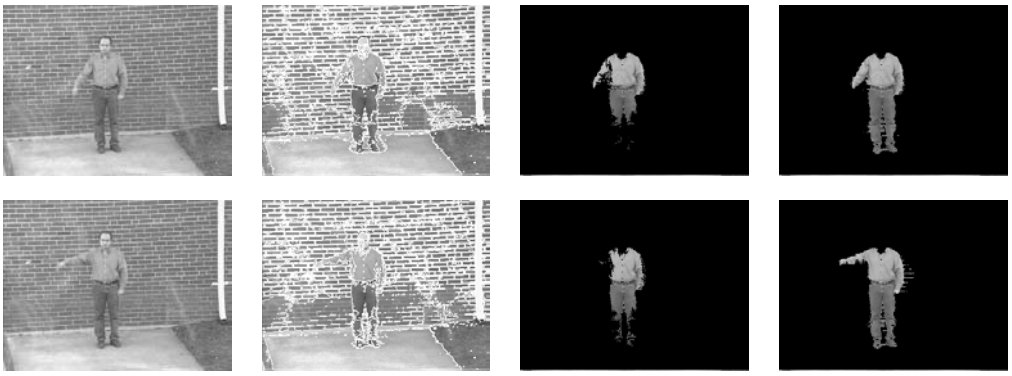
7.2.1 Combining Region Segmentation and Edge Continuity

Region segmentation may be coupled with edge continuity for object recognition. The preliminary concept is presented here. Given an image, color clustering [36] is employed to compute region segmentation - Figure 7.1 shows two images and the corresponding segmentation obtained. Next, the pose shown in Figure 7.2 is used to obtain segments belonging to the subject -shown in Figure 7.1(c). The obtained set is expanded to include other segments in the image that might belong to the subject - shown in Figure 7.1(d). This is done using the proposed edge affinity model. A lack of correspondence between the estimated silhouette of the person and the pose indicates a mismatch.

7.2.2 Regularization

The edge affinity model may be employed for gradient-dependent regularization. Two possible applications:

1. In active contours or level sets, the edge affinities may be used to apply a tangential stretching force on the active contour when it covers only a part of a contiguous curve in the image. See Figure 7.3 for an illustration.



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

Figure 7.1: Segmentation coupled with edge continuity for object recognition. (a) Images, (b) Segmentation, (c) Segments selected by reference pose, and (d) Expanded set of segments obtained by employing the proposed edge affinity model.

Figure 7.2: Reference pose used to select initial set of segments belonging to the subject.

2. A foreground-background separation approach proposed in [5] suppresses image gradients estimated to belong to the background. Edge affinities may be potentially used to regularize the suppression. If two edges have high affinity then either both or none should be suppressed.

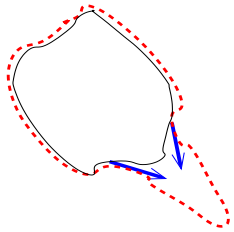


Figure 7.3: The edge affinities may be employed to exert a tangential stretching force on active contours to improve their convergence. The red-dashed plot is a curve on the image and the black-solid line is the active contour. If two pairs of edges have high affinity then either both or none should be aligned with the active contour. If only one of a pair of image edges is aligned with the active contour then it exerts a tangential force on the contour to stretch it onto its neighbor.

Bibliography

[1] MSU Graphics and Media Lab, Computer Vision Group, http://graphics.cs.msu.ru.

[2] OSU Support Vector Machines (SVMs) Toolbox, http://www.csie.ntu.edu.tw/c̃jlin/libsvm.

[3] EC Funded CAVIAR project/IST 2001 37540. Found at http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[4] CMU Graphics Lab Motion Capture Database, http://mocap.cs.cmu.edu, 2005.

[5] A. Agrawal, R. Raskar, and R. Chellappa. Edge suppression by gradient field transformation using cross-projection tensors. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR-2006)*, 2006.

[6] R. M. Alexander. Optimum timing of muscle activation of simple models of throwing. *Jnl. Theoretical Biology*, 150(3):349–372, 1991.

[7] O. Arikan, D. A. Forsyth, and J. F. O'Brien. Motion synthesis from annotations. *ACM Trans. Graph.*, 22(3):402–408, 2003.

[8] D. G. Asatryan and A. G. Fel'dman. Functional tuning of the nervous system with control of movement or maintenance of a steady posture. *Biophysics*, 1(10):925–935, 1965.

[9] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. PAMI*, 24(4):509–522, 2002.

[10] J. Besag. On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society, Series B*, 48:259–302, 1986.

[11] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. and Machine Intell.*, 23(3):257–267, Mar. 2001.

[12] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *Proc. IEEE Workshop on Perceptual Organization in Computer Vision, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR-2004)*, June, 2004.

[13] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV-2002*, volume 2, pages 109–124, 2002.

[14] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. and Machine Intell.*, (9), Sep. 2004.

[15] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. and Machine Intell.*, 20(11):1222–1239, Nov. 2001.

[16] C. Bregler. Learning and recognizing human dynamics in video sequences. In *CVPR-1997*, 1997.

[17] M. A. Butt and P. Maragos. Optimum design of chamfer distance transforms. *IEEE Trans. Image Proc.*, 7(10):1477–1484, Oct. 1998.

[18] P. B. Chou and C. M. Brown. The theory and practice of Bayesian image modelling. *Int'l J. Computer Vision*, 4:185–210, 1990.

[19] C.-C. Chu and J. K. Aggarwal. The integration of image segmentation maps using region and edge information. *IEEE Trans. Pattern Anal. and Machine Intell.*, 15(12):1241–1252, Dec. 1993.

[20] D. Cremers, N. A. Sochen, and C. Schnorr. Multiphase dynamic labeling for variational recognition-driven image segmentation. In *ECCV-2004*, volume 4, pages 74–86, 2004.

[21] R. Cross. Physics of overarm throwing. *Am. J. Physics*, 72(3):305–312, 2004.

[22] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR-2005*, 2005.

[23] J. H. Elder, A. Krupnik, and L. A. Johnston. Contour grouping with prior models. *IEEE Trans. Pattern Anal. and Machine Intell.*, 25(6):661–674, June 2003.

[24] A. Elgammal, V. D. Shet, Y. Yacoob, and L. S. Davis. Learning dynamics for exemplar-based gesture recognition. In *CVPR-2003*, volume 1, pages 571–578, June 18-20, 2003.

[25] P. Felzenszwalb. Learning models for object recgonition. In *CVPR-2001*, pages 1056–1062, 2001.

[26] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. In *CVPR'04*, 2004.

[27] M. Flanders, L. Daghestani, and A. Berthoz. Reaching beyond reach. *Exp. Brain Res.*, 126:19–30, 1999.

[28] T. Flash and N. Hogan. The coordination of arm movements: An experimentally confirmed mathematical model. *J. Neurosci.*, 5:1688–1703, Jul. 1985.

[29] T. Gautama and M. M. V. Hulle. A phase-based approach to the estimation of the optical flow field using spatial filtering. *IEEE Trans. Neural Net.*, 13(5):1127–1136, 2002.

[30] D. Gavrila. Multi-feature hierarchical template matching using distance transforms. In *ICPR-1998*, pages 439–444, 1999.

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

[32] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. and Machine Intell.*, 6:721–741, 1984.

[33] S. Geman and C. Graffigne. Markov random field image models and their application to computer vision. In *Int'l Cong. of Mathematicians*, pages 1496–1517, 1986.

[34] G. Guy and G. Medioni. Inferring global perceptual contour from local features. *Int'l J. Computer Vision*, 20(1-2):113–133, 1996.

[35] P. Haggard and J. Richardson. Spatial patterns in the control of human arm movements. *J. Exp. Pshychol. Human Percept. Performance*, 22:42–62, 1996.

[36] B. Heisele, U. Kressel, and W. Ritter. Tracking non-rigid, moving objects based on color cluster flow. In *Proc. 1997 IEEE Conf. Computer Vision and Pattern Recognition (CVPR '97)*, pages 257–261, 1997.

[37] B. Hoff and M. A. Arbib. Models of trajectory formation and temporal interaction of reach and grasp. *J. Motor Behavior*, 25(3):175–192, 1993.

[38] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR-2006)*, 2006.

[39] C. Huang, H. Ai, Y. Li, and S. Lao. Vector boosting for rotation invariant multi-view face detection. In *ICCV-2005*, pages 446–453, 2005.

[40] I. H. Jermyn and H. Ishikawa. Globally optimal regions and boundaries as minimum ratio weight cycles. *IEEE Trans. Pattern Anal. and Machine Intell.*, 23(10):1075–1088, 2001.

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

[42] G. Kanizsa. Subjective contours. *Sci. Am.*, 234:48–52, 1976.

[43] S. W. Keele. Behavioral analysis of movement. In V. B. Brooks, editor, *Handbook of Physiology, Section 1, Volume II, Part 2*. 1981.

[44] V. Kolmogorov and C. Rother. Comparison of energy minimization algorithms for highly connected graphs. In *ECCV'06*, 2006.

[45] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. and Machine Intell.*, 26(2):147–159, Feb. 2004.

[46] M. P. Kumar, P. Torr, and A. Zisserman. Obj cut. In *CVPR-2005*, 2005.

[47] M. W. Lee and I. Cohen. Proposal maps driven MCMC for estimating human body pose in static images. In *CVPR-2004*, volume 2, pages 334–341, 2004.

[48] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *CVPR-2003*, volume 2, pages 409–415, 2003.

[49] T. K. Leung and J. Malik. Contour continuity in region based image segmentation. In *Proc. European Conf. Computer Vision (ECCV'98)*, volume 1, pages 544–559, 1998.

[50] J. Luo and C. Guo. Perceptual grouping of segmented regions in color images. *Pattern Recognition*, 36:2781–2792, 2003.

[51] R. G. Marteniuk, C. L. MacKenzie, M. Jeannerod, S. Athenes, and C. Dugas. Constraints on human arm movement trajectories. *Canadian Jnl. Psychology*, 41(3):365–378, 1987.

[52] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detector. In *ECCV-2004*, pages 69–82, 2004.

[53] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, Nov. 2006.

[54] A. Mohan, B. Papageorgiou, and T. Poggio. Optimum design of chamfer distance transforms. *IEEE Trans. Pattern Anal. and Machine Intell.*, 20(4):349–361, Apr. 2001.

[55] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *Proc. European Conf. Computer Vision, (ECCV'02)*, volume 3, pages 666–680, 2002.

[56] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR-2004*, volume 2, pages 326–333, June 2004.

[57] C. F. Olson and D. P. Huttenlocher. Automatic target recognition by matching oriented edge pixels. *IEEE Trans. Image Proc.*, 6(1):103–113, Jan. 1997.

[58] P. Parent and S.W.Zucker. Trace inference, curvature consistency and curve detection. *IEEE Trans. Pattern Anal. and Machine Intell.*, 11(8):823–839, Aug. 1989.

[59] P. Peursum, G. West, and S. Venkatesh. Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In *ICCV-2005*, pages 82–89, 2005.

[60] P. Pigeon, S. B. Bortolami, P. Dizio, and J. R. Lackner. Coordinated turn and reach movements II: Planning in an external reference frame. *J. Neurophsio.*, 89:290–303, 2003.

[61] P.L.Rosin and G. West. Salience distance transforms. *Computer Vision, Graphics and Image Processing - Graphical Models and Image Processing*, 57(6):483–521, Nov. 1995.

[62] V. S. N. Prasad, L. S. Davis, S. D. Tran, and A. Elgammal. Edge affinity for pose-contour matching. *Computer Vision and Image Understanding*, 104(1):36–47, Oct. 2006.

[63] V. S. N. Prasad, V. Kellokompu, and L. S. Davis. Ballistic hand movements. In *Conf. Articulated Motion and Deformable Objects*, 2006.

[64] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. IEEE*, volume 77, pages 257–286, Feb. 1989.

[65] D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. In *CVPR-2003*, volume 2, pages 467–474, June 18-20, 2003.

[66] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR-2005*, 2005.

[67] C. Rao, M. Shah, and T. Syeda-Mahmood. Invariance in motion analysis of videos. In *MULTIMEDIA '03: Proc. 11th ACM Int'l Conf. Multimedia*, pages 518–527, 2003.

[68] L. Ren, A. Patrick, A. A. Efros, J. K. Hodgins, and J. M. Rehg. A data-driven approach to quantifying natural human motion. *ACM Trans. Graph.*, 24(3):1090–1097, 2005.

[69] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV-2003*, volume 1, pages 10–17, 2003.

[70] T. J. Roberts, S. J. McKenna, and I. W. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. In *ECCV-2004*, volume 4, pages 291–303, Apr. 2004.

[71] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *ECCV-2002*, 2002.

[72] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut" - interactive foreground extraction using iterative graph cuts. In *SIGGRAPH'04*, 2004.

[73] A. Sepehri, Y. Yacoob, and L. S. Davis. Parametric hand tracking for recognition of virtual drawings. In *Proc. Fourth IEEE Int'l Conf. Computer Vision Systems 2006 (ICVS'06)*, page 6, 2006.

[74] A. Sha'Ashua and S. Ullman. Structural saliency: The detection of globally salient structures using a locally connected network. In *ICCV-1988*, pages 321–327, 1988.

[75] E. Sharon, A. Brandt, and R. Basri. Segmentation and boundary detection using multiscale intensity measurements. In *CVPR-2001*, volume 1, pages 469–476, 2001.

[76] E. Shechtman and M. Irani. Space-time behavior based correlation. In *CVPR-2005*, pages 405–412, 2005.

[77] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the space of human actions. In *ICCV-2005*, volume 1, pages 144–149, 2005.

[78] V. D. Shet, V. S. N. Prasad, A. Elgammal, Y. Yacoob, and L. S. Davis. Multi-cue exemplar-based nonparametric model for gesture recognition. In *Proc. Indian Conf. Computer Vision, Graphics and Image Processing (ICVGIP) 2004*, 16-18 Dec., 2004.

[79] I. Smyth and M. Wing, editors. *The Psychology of Human Movement*. Academic Press Inc., Orlando, FL 32887, 1984.

[80] J. F. Soechting, C. A. Buneo, U. Herrmann, and M. Flanders. Moving effortlessly in three dimensions: Does Donder's Law apply to arm movement? *J. Neurosci.*, 15:6271–6280, 1995.

[81] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields. In *ECCV'06*, 2006.

[82] A. Thayananthan, B. Stenger, P. Torr, and R. Cipolla. Shape context and chamfer matching in cluttered scenes. In *CVPR-2003*, volume 1, pages 127–133, June 18-20, 2003.

[83] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Proc. IEEE Int'l Conf. Computer Vision (ICCV-2003)*, 2003.

[84] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Proc. Int'l Conf. Computer Vision (ICCV'01)*, pages 50–59, 2001.

[85] Y. Uno, M. Kawato, and R. Suzuki. Formation and control of optimal trajectory in human multijoint arm movement. *Biol. Cybernetics*, pages 89–101, 1989.

[86] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR-2001*, volume 1, pages 511–518, 2001.

[87] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *ICCV-2003*, 2003.

[88] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. MAP-estimation via agreement on (hyper)trees - message passing and linear programming approaches. *IEEE Trans. Inform. Theory*, 51, 2005.

[89] D. Weinland, R. Ronfard, and E. Boyer. Automatic discovery of action taxonomies from multiple views. In *CVPR-2006*, pages 1639–1645, 2006.

[90] L. R. Williams and D. W. Jacobs. Local parallel computation of stochastic completion fields. *Neural Computation*, 9:859–881, 1997.

[91] A. D. Wilson and A. F. Bobick. Parametric hidden markov models for gesture recognition. *IEEE Trans. Pattern Anal. and Machine Intell.*, 21(9):884–900, Sep. 1999.

[92] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In *ICCV'05*.

[93] B. Wu and R. Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *CVPR'06*.

[94] A. Yilmaz and M. Shah. Actions as objects: A nover action representation. In *CVPR-2005*, 2005.

[95] A. Yilmaz and M. Shah. Recognizing human action in videos acquired by uncaliberated moving cameras. In *ICCV-2005*, 2005.

[96] S. Yu and J. Shi. Perceiving shapes through region and boundary interaction. Technical Report CMU-RI-TR-01-21, Robotics Institute, Carnegie Mellon Univ., Pittsburgh, PA, July 2001.

[97] S. X. Yu, R. Gross, and J. Shi. Concurrent object recognition and segmentation by graph partitioning. In *Neural Information Processing Systems*, pages 1383–1390, 3-8 Dec., 2001.

[98] J. Zhang, R. Collins, and Y. Liu. Representation and matching of articulated shapes. In *CVPR-2004*, volume 2, pages 342–349, June 2004.

[99] Q. Zhu, S. Avidan, M.-C. Yeh, and K.-T. C. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR'06*.