M.I.T. Media Lab Perceptual Computing Group Technical Report No. 387

Appearance-Based Motion Recognition of Human Actions

by
James William Davis

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on July 25, 1996
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

Abstract

A new view-based approach to the representation and recognition of action is presented. The work is motivated by the observation that a human observer can easily and instantly recognize action in extremely low resolution imagery with no strong features or information about the three-dimensional structure of the scene. Our underlying representations for action are view-based descriptions of the coarse image motion. Using these descriptions, we propose an appearance-based recognition strategy embedded within a hypothesize-and-test paradigm.

A *binary motion region* (BMR) image is initially computed to act as an index into the action library. The BMR grossly describes the spatial distribution of motion energy for a given view of a given action. Any stored BMRs that plausibly match the unknown input BMR are then tested for a coarse, categorical agreement with a known motion model of the action.

We have developed two motion-based methods for the verification of the hypothesized actions. The first approach collapses the temporal variations of region-based motion parameters into a single, low-order coefficient vector. A statistical acceptance region generated around the coefficients is used for classification into the training instances. In the second approach, a *motion history image* (MHI) is the basis of the representation. The MHI is a static image where pixel intensity is a function of the recency of motion in a sequence. Recognition is accomplished in a feature-based statistical framework.

Results employing multiple cameras show reasonable recognition within a MHI verification method which automatically performs temporal segmentation, is invariant to linear changes in speed, and runs in real-time on a standard platform.

Thesis Supervisor: Aaron F. Bobick
Title: Assistant Professor of Computational Vision

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recently in computer vision, a growing interest in video sequences, rather than single images, has emerged. Much focus is currently being placed on the understanding and interpretation of "action" (e.g. sitting, throwing, walking, etc). Understanding action is particularly attractive to those developing wireless interfaces [12, 13, 14] and interactive environments [8]. Such systems need to recognize specific actions of the person (e.g. hand gestures, body movements) while not appearing intrusive to the user. Having the user "suit-up" with positional sensors or color-marked clothing may inhibit movement of the user, making the experience feel "un-natural." Additionally, there may be a constraint present requiring the person to be positioned at a certain location and orientation in the field-of-view. Permitting the user to freely perform actions from a variety of views without any special body-tracking suit would be a significant advancement in the design of systems that analyze human movements.

Rather than exploiting the temporal motion of actions, many approaches use specific body poses (e.g. "arm out from side") for representing actions (e.g. "reaching"). The underlying assumption here is that a collection of static frames of the body contains the necessary information of what action is performed. But to see someone "reaching", for example, is it important to analyze the *entire* body? Consider a person waving hello to another person. It shouldn't matter if the person performing the waving action is standing or sitting, and these two cases most likely should not be modeled separately. It does not seem appropriate to examine the entire body for actions that don't include full-body activity. An alternative would be to pay attention to the *motion* occurring. The motion regions of the

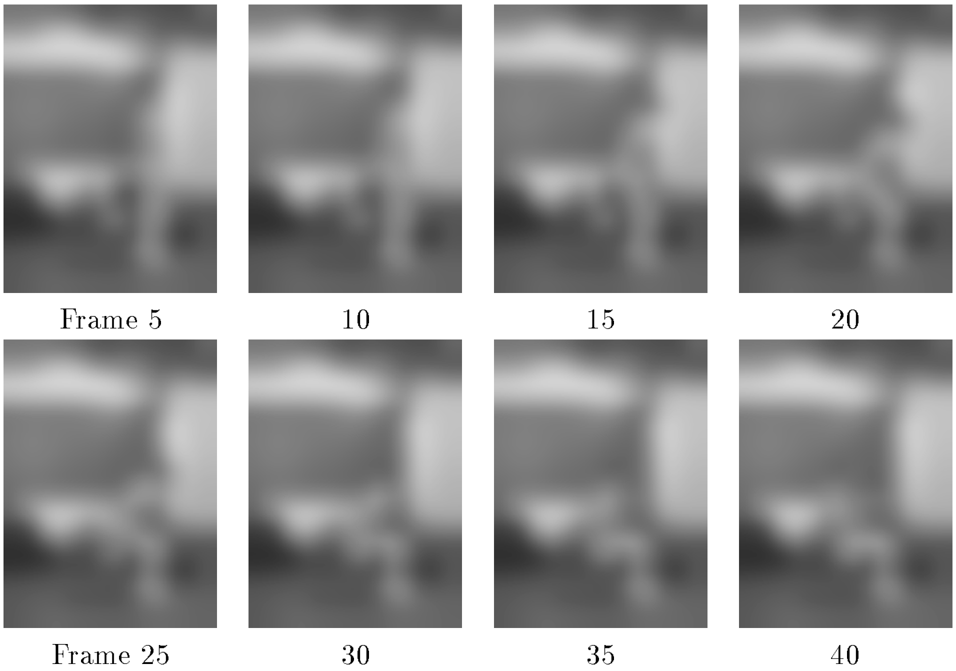


Figure 1-1: Selected frames from video of someone performing an action. Almost no structure is present in each frame, nor are there any features from which to compute a structural description. Yet people can trivially recognize the action as someone sitting.

body could be used to drive the attention of a modeling mechanism to examine particular limbs, for example. More fundamentally, the motion pattern itself, as opposed to the specific body pose(s), could be used for recognition.

In this thesis we develop two view-based approaches to the representation and recognition of action using motion characteristics. We consider the motion of the action to be a salient indicator for recognition of various body movements. In the following section we discuss our motivation for such an approach and then describe the method in more detail.

1.1 Motivation

The motivation for the approach presented in this thesis can be demonstrated in a single video-sequence. Unfortunately the media upon which these words are printed precludes the reader from experiencing the full impact of viewing the video. A poor substitute is a collection of selected frames of the video, as shown in Figure 1-1.

The video is a significantly blurred sequence — in this case an up-sampling from images of resolution 15x20 pixels — of a human performing a simple, yet readily recognizable, activity — sitting. Most people shown this video can identify the action in less than one second from the start of the sequence. What should be quite apparent is that most of the individual frames contain no discernible image of a human being. Even if a system knew that the images were that of a person, no particular pose could be reasonably assigned due to the lack of features present in the imagery.

A more subtle observation is that no good features exist upon which to base a structure-from-motion algorithm [32]. This distinction is important. Although individual frames of moving light displays also contain insufficient information to directly recover pose, they *do* contain features that allow for the structural recovery of the limbs [18] without *a priori* knowledge of the semantic assignments (e.g. "light 1 is the left hip"). One may not be able to prove that the blurred sequence of Figure 1-1 cannot be analyzed for three-dimensional articulated structures before the assignment of body parts to image regions, however, the lack of any image detail (good trackable features) makes such a possibility remote in many of the current 3-D model-tracking methods; it would seem that an initial alignment of the model to the image is necessary. Many approaches have been proposed with the presumption that 3-D information would be useful and perhaps even necessary to understand actions (e.g. [28, 16, 19, 4, 29, 15]), but the blurred sequence example leads us to believe that much information about the action is present in the underlying motion of the person.

When viewing the motion in a blurred sequence, two distinct patterns are apparent. The first is the spatial region in which the motion is occurring. The pattern is defined by the area of pixels *where* something is changing largely independent of how it is moving. The second pattern is *how* the motion itself is behaving within these regions (e.g. an expanding or rotating field in a particular location). We developed our methods to exploit these notions of *where* and *how*, believing that these observations capture significant motion properties of actions that can be used for recognition.

It should be noted that the ideas expressed in this thesis concerning motion recognition are based on the fact that humans *can* recognize movements in blurred motion sequences. The work is *not* attempting to emulate the human visual system.

1.2 Applications

A system capable of recognizing human actions would be useful to wide variety of applications. Many entertainment systems (e.g. video/arcade games) could be made more compelling using an action recognition system, where the users control navigation by using their own body movements. Having the user control the graphic character's run, jump, kick, and punch actions using the corresponding true body motion would give the player a deeper feeling of involvement in the game. Similarly, a virtual aerobics instructor could be developed in this framework that watches the exercise movements of individuals and offers feedback on the performance of the exercises[1]. For interactive environments, allowing for natural and unencumbered movements provides a more realistic and enjoyable experience for the user. Also, many monitoring situations would be improved if a system were able to detect particular behaviors which are considered harmful or dangerous, as opposed to just any movement (e.g. consider a baby monitoring system which notifies the parent only when a potentially harmful situation arises rather than every time the baby moves). All of the above applications require the ability to automatically and unobtrusively recognize human actions.

1.3 Approach

Given that motion recognition is possible in the absence of trackable features (as shown in the blurred sequence), a question arises. How might the recognition be accomplished? For us, the most straightforward answer is that the motion pattern itself is to be recognized. It should be possible to recognize a two-dimensional motion pattern as an instance of a motion field that is consistent with how some known movement appears when viewed from a particular direction. Since these motion patterns are different for various views of actions, a view-based, model-based technique is required. The model, however, is of the body's *motion* and *not* of the body's configuration.

The basic components of the theory presented are embedded in a hypothesize-and-test paradigm [17] to eliminate exhaustive searching for possible matches. First, a simple feature-based characterization of the spatial ("where") motion region is used as the initial filter into

[1] In this work, we have begun to embrace aerobics, including many exercise movements in some of the testing procedures. We plan on working more extensively within the aerobic domain in the future.

the set of known action models. Then two different methods representing the directional ("how") motion are presented which find a match, if one exists, with the hypothesized models. The first method reduces motion time-traces of particular regions, or patches, of the image down to a single low-order coefficient vector. Statistical recognition on the coefficient vectors is used for classification into the training instances. The second method is a real-time approach which collapses the temporal motion information of the action sequence into a single image. A simple feature set is extracted from the image and used to match against the known movements. These methods demonstrate the plausibility of using motion, as opposed to using 3-D reconstruction or body poses, as a means of recognizing action.

1.4 Outline of thesis

The remainder of this thesis examines our approach to action representation and recognition. In the next chapter, the context of this work with related research in two relevant areas is examined (configuration-based tracking and recognition and motion-based recognition). Chapter 3 discusses the hypothesis generation used to reduce the number of action models to be examined for verification. We use a measure of the spatial distribution of motion as a course index into the action library. Chapter 4 describes two motion model methods developed for action representation and recognition. First, a parameterized motion model that describes regional motion characteristics over time is examined. Second, a static representation of motion is presented which collapses the motion information of a sequence into a single image. Next, the results of a real-time implementation of the second method are given in Chapter 5. Lastly, Chapter 6 summarizes the work presented and discusses future extensions.

Chapter 2

Previous work

The number of approaches to recognizing motion and action has recently grown at a tremendous rate. For an excellent survey on the machine understanding of motion (particularly human motion) see the work of Cédras and Shah [5]. A detailed description of methods for extracting motion information (e.g. optical flow, motion correspondence, trajectory parameterization) and a brief discussion on matching is reviewed. Also, recent work in motion recognition in terms of generic motion (trajectory and cyclic), human movements, and specialized motion (lip-reading, gestures) is discussed. We divide the prior work that inspired the approach presented in this thesis into two general areas: configuration-based tracking and recognition and motion-based recognition.

2.1 Configuration-based tracking and recognition of action

2.1.1 Tracking

The first and most obvious body of relevant work includes the approaches using structural or appearance-based representations to tracking and understanding human action. Some believe that a 3-D description is necessary and sufficient for understanding action (e.g. [19, 4, 29, 15, 28, 16]), while others choose to analyze the 2-D appearance as a means of interpretation (e.g. [6, 7, 1, 35]). We now take a closer look at these approaches.

The most common method for attaining the 3-D information in the action is to recover the pose of the object at each time instant using a 3-D model of the object. A common method for model fitting in these works is to use a residual measure between the projected model and object contours (e.g. edges of body in the image). This generally requires a strong segmentation of foreground/background and also of the individual body parts to aid the model alignment process. It is difficult to imagine such techniques could be extended to the blurred sequence of Figure 1-1.

For example, Rehg and Kanade [28] used a 27 degree-of-freedom (DOF) model of a human hand in their system called "Digiteyes". Local image-based trackers are employed to align the projected model lines to the finger edges against a solid background. The work of Goncalves et al. [16] promoted 3-D tracking of the human arm against a uniform background using a two cone arm model and a single camera. Though it may be possible to extend their approach to the whole body as claimed, it seems unlikely that it is appropriate for non-constrained human motion with self-occlusion. Hogg [19] and Rohr [29] used a full-body cylindrical model for tracking walking humans in natural scenes. Rohr incorporates a 1 DOF pose parameter to aid in the model fitting. All the poses in a walking action are indexed by a single number. Here there is only a small subset of poses which can exist. Gavrila and Davis [15] also used a full-body model (22 DOF, tapered super-quadrics) for tracking human motion against a complex background. For simplifying the edge detection in cases of self-occlusion, the user is required to wear a tight-fitting body suit with contrasting limb colors.

One advantage of having the recovered model is the ability to estimate and predict the feature locations, for instance edges, in the following frames. Given the past history of the model configurations, prediction is commonly attained using Kalman filtering [29, 28, 16] and velocity constraints [26, 15].

Because of the self-occlusions that frequently occur in articulated objects, some employ multiple cameras and restrict the motion to small regions [28, 15] to help with projective model occlusion constraints. A single camera is used in [19, 16, 29], but the actions tracked in these works had little deviation in the depth of motion. Acquiring the 3-D information from image sequences is currently a complicated process, many times necessitating human intervention or contrived imaging environments.

2.1.2 Recognition

As for action recognition, Campbell and Bobick [4] used a commercially available system to obtain 3-D data of human body limb positions. Their system removes redundancies that exist for particular actions and performs recognition using only the information that varies between actions. This method examines the relevant parts of the body, as opposed to the entire body data. Siskind [31] similarly used known object configurations. The input to his system consisted of line-drawings of a person, table, and ball. The positions, orientations, shapes, and sizes of the objects are known at all times. The approach uses support, contact, and attachment primitives and event logic to determine the actions of dropping, throwing, picking up, and putting down. These two approaches address the problem of recognizing actions when the precise configuration of the person and environment is known while the methods from the previous section concentrate on the recovery of the object pose.

In contrast to the 3-D reconstruction and recognition approaches, others attempt to use only the 2-D appearance of the action (e.g. [1, 7, 6, 35]). View-based representations of 2-D statics are used in a multitude of frameworks, where an action is described by a sequence of 2-D instances/poses of the object. Many methods require a normalized image of the object (usually with no background) for representation. For example, Cui et al. [6], Darrell and Pentland [7], and also Wilson and Bobick [33] present results using actions (mostly hand gestures), where the actual grayscale images (with no background) are used in the representation for the action. Though hand appearances remain fairly similar over a wide range of people, with the obvious exception of skin pigmentation, actions that include the appearance of the total body are not as visually consistent across different people due to obvious natural variations and different clothing. As opposed to using the actual raw grayscale image, Yamato et al. [35] examines body silhouettes, and Akita [1] employs body contours/edges. Yamato utilizes low-level silhouettes of human actions in a Hidden Markov Model (HMM) framework, where binary silhouettes of background-subtracted images are vector quantized and used as input to the HMMs. In Akita's work [1], the use of edges and some simple 2-D body configuration knowledge (e.g. the arm is a protrusion out from the torso) are used to determine the body parts in a hierarchical manner (first find legs, then head, arms, trunk) based on stability. Individual parts are found by chaining local contour information. These two approaches help alleviate *some* of the variability between people but introduce other problems such as the disappearance of movement that happens to be within the silhouetted region and also the varying amount of contour/edge information that arises when the background or clothing is high versus low frequency (as in most natural scenes). Also, the problem of examining the entire body, as opposed to only the desired regions, still exists, as it does in much of the 3-D work.

Whether using 2-D or 3-D structural information, many of the approaches discussed so far consider an action to be comprised of a sequence of poses of an object. Underlying all of these techniques is the requirement that there be individual features or properties that can be extracted and tracked from each frame of the image sequence. Hence, motion understanding is really accomplished by recognizing a sequence of static configurations. This understanding generally requires previous recognition and segmentation of the person [26]. We now consider recognition of action within a motion-based framework.

2.2 Motion-based recognition

Directional motion recognition [26, 30, 24, 3, 34, 31, 11] approaches attempt to characterize the motion itself without reference to the underlying static poses of the body. Two main approaches include the analysis of the body region as a single "blob-like" entity and the tracking of predefined *regions* (e.g. legs, head, mouth) using motion instead of structural features.

Of the "blob-analysis" approaches, the work of Polana and Nelson [26], Shavit and Jepson [30], and also Little and Boyd [24] are most applicable. Polana and Nelson use repetitive motion as a strong cue to recognize cyclic walking motions. They track and recognize people walking in outdoor scenes by gathering a feature vector, over the entire body, of low-level motion characteristics (optical-flow magnitudes) and periodicity measurements. After gathering training samples, recognition is performed using a nearest centroid algorithm. By assuming a fixed height and velocity of each person, they show how their approach is extendible to tracking multiple people in simple cases. Shavit and Jepson also take an approach using the gross overall motion of the person. The body, an animated silhouette figure, is coarsely modeled as an ellipsoid. Optical flow measurements are used to help create a phase portrait for the system, which is then analyzed for the force, rotation, and strain dynamics. Similarly, Little and Boyd recognize people walking by analyzing the motion associated with two ellipsoids fit to the body. One ellipsoid is fit using the motion region silhouette of the person, and the other ellipsoid is fit using motion magnitudes as weighting factors. The relative phase of various measures (e.g. centroid movement, weighted centroid movment, torque) over time for each of the ellipses characterizes the gait of several people.

There is a group of work which focuses on motions associated with facial expressions (e.g. characteristic motion of the mouth, eyes, and eyebrows) using region-based motion properties [34, 3, 11]. The goal of this research is to recognize human facial expressions as a dynamic system, where the motion of interest regions (locations known *a priori*) is relevant. Their approaches characterize the expressions using the underlying motion properties rather than represent the action as a sequence of poses or configurations. For Black and Yacoob [3], and also Yacoob and Davis [34], optical flow measurements are used to help track predefined polygonal patches placed on interest regions (e.g. mouth). The parameterization and location relative to the face of each patch was given *a priori*. The temporal trajectories of the motion parameters were qualitatively described according to positive or negative intervals. Then these qualitative labels were used in a rule-based, temporal model for recognition to determine emotions such as anger or happiness. Recently, Ju, Black, and Yacoob [22] have extended this work with faces to include tracking the legs of a person walking. As opposed to the simple, independent patches used for faces, an articulated three-patch model was needed for tracking the legs. Many problems, such as large motions, occlusions, and shadows, make motion estimation in that situation more challenging than for the facial case. We extend this expression recognition approach in our work by applying a similar framework to the domain of full-body motion.

Optical flow, rather than patches, was used by Essa [11] to estimate muscle activation on a detailed, physically-based model of the face. One recognition approach classifies expressions by a similarity measure to the typical patterns of muscle activation. Another recognition method matches motion energy templates derived from the muscle activations. These templates compress the activity sequence into a single entity. In our second motion modeling method, we develop similar templates, but our templates represent the temporal motion characteristics rather than the overall energy.

2.3 Summary

Two main research areas closely related to the work presented in this thesis are configuration-based tracking and recognition and motion-based recognition.

We first examined several 2-D and 3-D methods for tracking and recognizing human movement using a sequence of static representations. We discussed approaches in which a 3-D description is stated to be necessary and sufficient for understanding action and others where the 2-D appearance is used as a means of interpretation. Next, motion recognition based on directional information was addressed. These approaches attempt to characterize the motion itself without any reference to the underlying static poses of the body. We seek to extend the parameterized approach in our first motion modeling method by developing a framework incorporating multiple models and performing statistical recognition on the motion trajectories. In our second approach, we generate motion-based templates (similar to [11]) for representing actions.

Chapter 3

Spatial distribution of motion

Given a rich vocabulary of motions that are recognizable, an exhaustive matching search is not feasible, especially if real-time performance is desired. In keeping with the hypothesize-and-test paradigm [17], the first step is to construct an initial index into the known motion library. Calculating the index requires a data-driven, bottom up computation that can suggest a small number of plausible motions to test further. We develop a representation of the spatial distribution of motion (the *where*), which is independent of the form of motion (the *how*), to serve as our initial index.

3.1 Binary motion region (BMR)

Consider the example of someone sitting, as shown in Figure 3-1. The top row contains key frames from a sitting sequence (the full-resolution version corresponding to the blurred sequence in Figure 1-1). The bottom row displays a cumulative *binary motion region* (BMR) image sequence corresponding to the frames above. The BMRs highlight regions in the image where any form of motion was present since the start of the action. As expected, the BMR sequence sweeps out a particular (and perhaps distinctive) region of the image. Our claim is that the shape of the region can be used to *suggest* both the action occurring and the viewing angle. Recognition cannot generally be performed with this representation only because certain motions may have very similar BMRs but possess different directional motion (e.g. "sitting down in a chair" versus "getting up out of the chair").

An obvious approach for constructing a BMR image is to first compute the optic flow field between each pair of frames using a local, gradient-based technique similar to Lucas

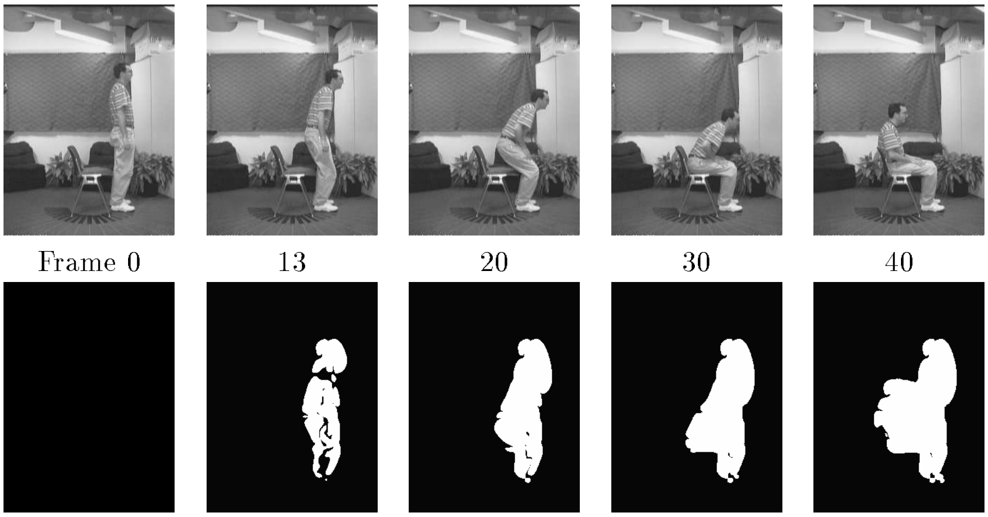


Figure 3-1: Example of someone sitting. Top row is keys frames. Bottom row is cumulative binary motion region images starting from Frame 0.

and Kanade [2] yielding a vector image $\vec{D}(x, y, t)$ for each sequential pair at time t. Then the BMR image is defined by

$$BMR(x, y, t) = \bigcup_{i=0}^{\tau-1} D'(x, y, t-i)$$

where $D'(x, y, t)$ is a binarized, thresholded version of the magnitude of $\vec{D}(x, y, t)$ designed to prevent noise in the motion computation from corrupting the process. The time duration parameter τ is controlled in the recognition procedure, which will be described later. However, the summation of the square of consecutive image differences (the method used here) often provides a more robust spatial motion-distribution signal. This is because image differencing determines only that pixels have changed, as opposed to where pixels have moved. Image differencing also permits real-time acquisition of the BMRs.

3.2 Motion-shape description using BMRs

Since the intent is for the BMR to capture the spatial distribution of motion, we selected a shape description vector to represent the BMR for comparison. Because the BMRs are blob-like in appearance, it seems reasonable to employ a set of moments-based descriptions for the characterization. The first seven parameters $< \nu_1, \ldots, \nu_7 >$ are the Hu moments

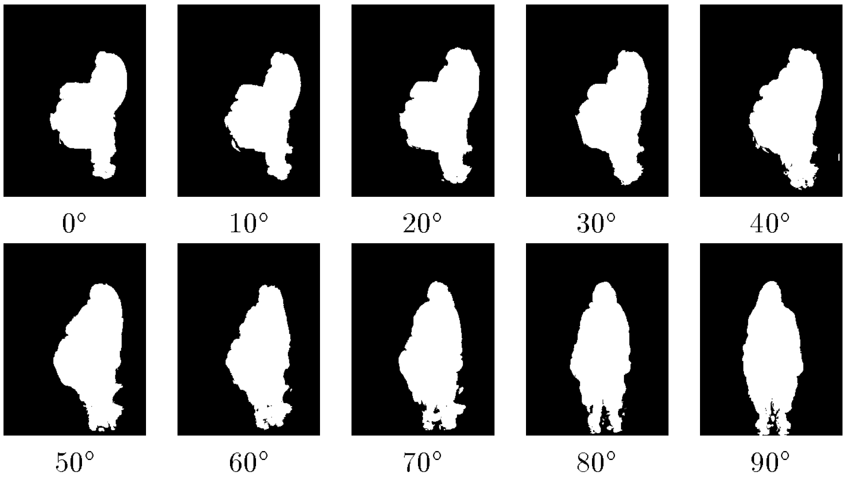


Figure 3-2: BMRs (aligned, averaged, and thresholded) of sitting action over a 90° viewing arc.

[20] (See Appendix A.3) which are known to yield reasonable shape discrimination in a translation, scale, and rotation invariant manner. We augment the feature vector to include terms sensitive to orientation and the correlation between the x and y locations: $\nu_8 = [E(xy) - E(x)E(y)]/[\sigma_x \sigma_y]$. Also, we include a measure of compactness ν_9 (computed as the ratio of the area of the image to the area of the best fit ellipse whose axes' orientation and relative size are determined by the principal components, and whose overall scale is set to a multiple of the standard deviation of the spatial distribution of the pixels in the BMR).

To illustrate the effectiveness of the BMR characterization, we performed the following experiment. First, we generated BMRs for several sitting actions. The complete data suite was 10 different viewing angles [0° (side view) through 90° (frontal view) in 10° increments] for 4 people each sitting twice, each time in a different chair. Then, an average BMR was computed for each angle by aligning the individual BMRs and averaging their intensity values[1]. Finally, the averaged result was thresholded. The result is shown in Figure 3-2.

Next we added 20 different aerobic exercises, shown in Figure 3-3, to the sitting data. The set includes a simple squat which in many ways is similar to sitting. The aerobics data set consists of 7 views (0° through 180° in 30° increments) of each of the 20 movements, with each sequence containing on the order of 100 frames. The corresponding BMRs for the aerobics data is shown in Figure 3-4. Therefore, with 140 aerobic BMRs and 10 sitting BMRs, the total target set consists of 150 BMRs.

[1] The alignment procedure used the motion region's principal components to calculate the translation and scale parameters necessary for alignment to some canonical position and scale

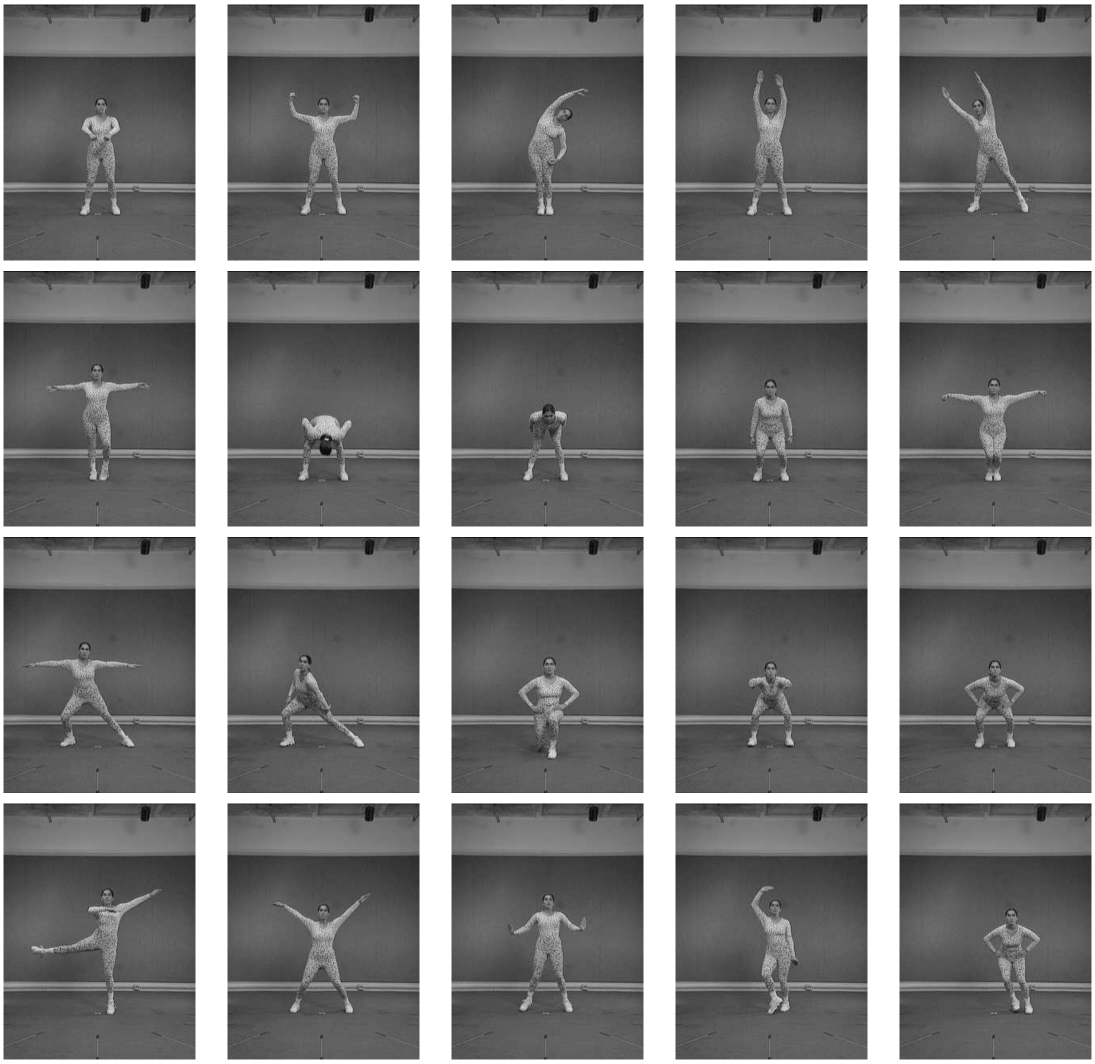


Figure 3-3: Example aerobic exercise movements.

The experiment consisted of testing sitting examples of a new subject against the entire training set of sitting and aerobics to find the most similar images. This test sitting data consisted of 2 repetitions for each view from $0°$ to $90°$ (in $10°$ increments) for a total of 20 inputs. For each of the calculated 20 input moment sets, the target BMR moment sets were ranked according to nearness using a metric of independent Mahalanobis distance[2], where a ranking of 1 is the closest element. The results are as follows: For the 20 input examples, the average rank of the correct sitting example (same angle as input) is 4.0, with a median of 3. This implies that typically the third or fourth best BMR out of 150 would in fact be

[2]That is, a distance with a diagonal Σ based upon all the data.

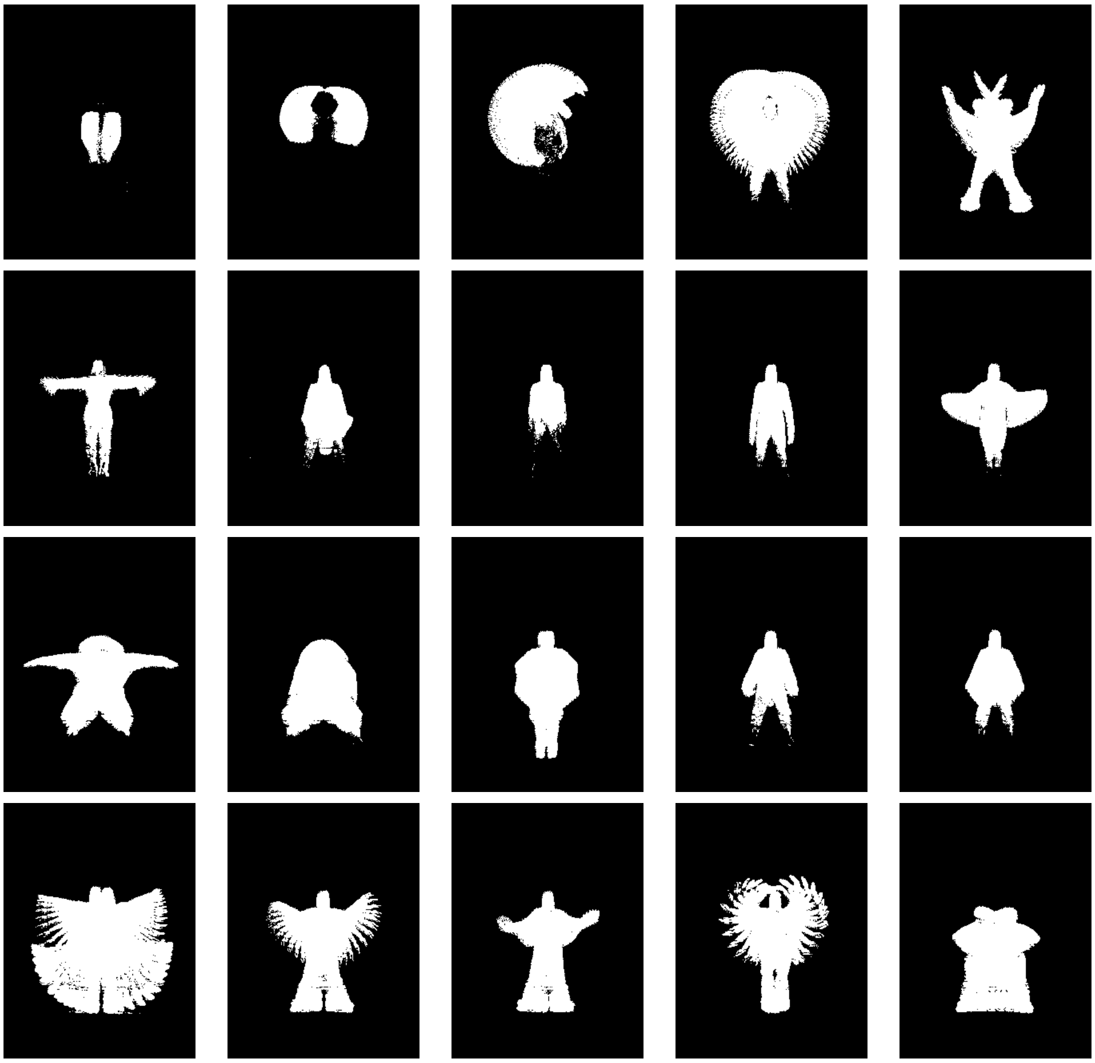


Figure 3-4: Aerobic exercise BMRs.

the correct match. Also, if there is some latitude in the verification procedure, then one only needs to find the correct action at a near-by viewing direction. The average ranking of the closest sitting BMR that was within 10° of the input move is 1.8, with a median of 1.5, and a worst case of 3. To find a close viewing direction of the correct action, typically only 1 or 2 of the 150 action-angle hypotheses need to be considered further for a match. These results lead us to believe that the BMRs would be a good index for a hypothesize-and-test paradigm where several hypotheses are considered.

3.3 BMR feature space

We now consider a BMR feature space which captures the relative statistics of the moment features. Instead of using a pooled covariance for computing the Mahalanobis distance, we compute a mean and covariance for the moments of each view of each action. We found that using only seven translation- and scale-invariant moments (x^p, y^q with order $p + q = 2, 3$) [20] (See Appendix A.4) offered reasonable discrimination. With these seven moments, the action performed can be translated in the image and placed at various depths (scale) from the novel position while retaining the same moment set (assuming no major camera distortions). These moments are simple to compute and require only a single pass of the image, thus making them feasible for real-time performance.

During training, many example BMR moment sets are gathered for each action, and a statistical model (mean, covariance) is calculated in a 7-D space for each action. For finding a match between a test moment set and a training action, we still use the Mahalanobis distance, but now using the action moment mean and covariance statistics. A full-covariance matrix captures the covarying moment information, rather than assuming independence (as in a diagonal covariance). Then, if the Mahalanobis distance between the unknown input and some example action is small, then that example is labeled as a *possible* classification for the unknown action and is included in the set of first-round plausible matches. In the following sections, we develop two verification procedures which use compact descriptions of the directional motion pattern.

Chapter 4

Motion models for representation and recognition of action

Given that a small subset of actions have been targeted by the BMR index, we must have a representation of the motion field which can be used for classifying an unknown movement as one of the hypothesized actions. We have developed two such approaches for representing actions which attempt to capture the overall regional motion. Our first approach uses specific region-based motion parameterizations for different viewing angles of various movements. The motion time-traces of particular regions ("patches") of the image are reduced down to a single, low-order coefficient vector. Statistical recognition on the coefficient vectors is used for classification. The second method is a real-time approach which collapses the motion information of the action sequence into a single image. A simple feature set is extracted from the image and used to match against the known movements. These methods demonstrate the plausibility of using motion as a means of recognizing action.

4.1 Motion Model 1: Region-based motion parameterization

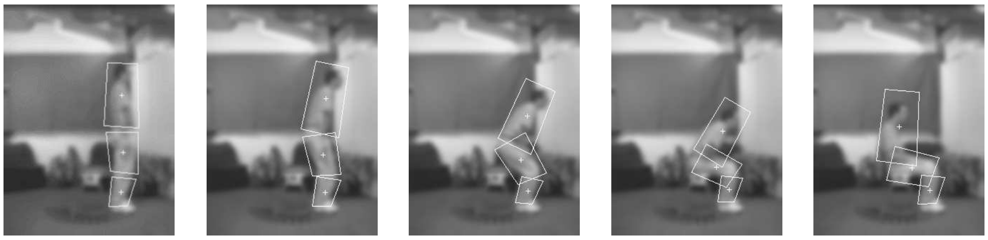
Our work in this section seeks to extend the facial-expression recognition work of Black and Yacoob [3]. In their work, the temporal trajectories of region-based motion parameters were qualitatively labeled to recognize facially expressive emotions (e.g. anger, happiness). They use only a single motion-tracking model consisting of five polygonal patches corresponding to the eyebrows, eyes, and mouth. This section seeks to extend that approach by allowing different motion models for various viewing angles of multiple movements. Clearly, the relevant motion fields to see a person "sitting" are not the same as those expected for seeing someone "walking". There is no reason to believe that the same parameterization of motion (or configuration of patches) would apply to both situations. We therefore construct a motion model which contains the motion parameterization necessary to describe the motion for a given action viewed from a given angle. The motion parameter traces resulting from these motion-patch models tracked over time are used to validate whether the motion field is consistent with a known view of a given action.

4.1.1 Parameterized motion representation

In this section we derive a motion description mechanism that collects and collapses region-based motion traces to a single, low-order vector. Then, for each view angle condition of each move acquired from training we can define an acceptance region as a probability distribution on that vector. Thus, if the vector for an unknown motion falls within the acceptance region for a given action, the unknown action is classified as an example of that action.

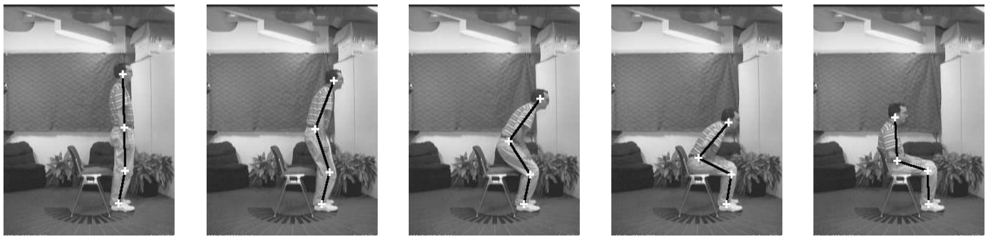
Our ultimate goal for this method is to have a set of polygonal patches whose placement and tracking are determined by the hypothesized action. The motion parameters would be determined by tracking the patches using a region-based parametric optic flow algorithm. One example of tracked patches is shown in Figure 4-1. Three polygonal patches were created and placed manually but tracked automatically using an affine model of optic flow [2]. We have not yet achieved a robust enough model placement and tracking algorithm to test our recognition method using patches. Unlike the face images of [3], full-body action sequences can have quite a variety of image textures, shadows, and occlusions which make motion estimation a non-trivial operation. Recent work by Ju, Black, and Yacoob [22] further investigate this patch tracking paradigm applied to the human body and have found similar shortcomings.

To decouple the nature of the representation from our current ability to do patch tracking, we employ a simplified patch model, which uses manually placed and tracked sticks (See Figure 4-2). A stick is defined by its two endpoints $\{<x_1, y_1>, <x_2, y_2>\}$, and



Frame 0 13 20 30 40

Figure 4-1: Automatic tracking of patches. The patches are tracked using a parametric (affine) optic-flow algorithm. The images are blurred to aid in optical-flow estimation.



Frame 0 13 20 30 40

Figure 4-2: Manual tracking of sticks. Sticks are manually placed and tracked on the torso, upper leg, and lower leg.

therefore has 4 degrees of freedom. As such, we can describe the motion of a stick from one frame to another with four numbers. To help maintain our intuitions, we will consider the four variations to be Trans-x, Trans-y, Rotation, and Scale. We relate them in the usual way to the algebraic transformation:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_1 & -a_2 \\ a_2 & a_1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a_3 \\ a_4 \end{bmatrix}$$

where $[x, y]^T$ is the position of an endpoint in one frame, $[x', y']^T$ is the position of the endpoint in the other frame, Trans-x = a_3, Trans-y = a_4, Scale = a_1, and Rotation = a_2. For the sitting example, as illustrated in Figure 4-2, we used three sticks. We note that it is unlikely that a single stick/patch representation can be robustly tracked over all possible views of an action (though three sticks are sufficient for sitting within a 180° view arc). Thus, the region parameterization of the motion description itself is generally sensitive to view angle. We refer to the space of view angles over which a region parameterization applies as a *motion aspect.* Here, a particular model attempts to span as wide an angular range as possible using a single, low order representation of the appearance of the motion. However, when this model is not applicable, a new model is created to accommodate this new discrete region (or aspect). Fortunately, the hypothesize-and-test method can be easily modified to use different stick/patch models for different views if necessary. The appropriate stick/patch model for tracking can be retrieved given the correct BMR hypothesis.

If we relate the sticks at each time t back to the original stick configurations at time $t = 0$, we can characterize the motion of the 3 sticks (for example) by a 12-dimensional (3 sticks $\times$ 4 parameters), time-dependent vector $\vec{M}(t)$. Scale invariance can be achieved by normalizing the Trans-x and Trans-x parameters with respect to the height of the person (top to bottom of stick model). For each given viewing direction α we would get a different motion appearance so we need to index the motion by α: $\vec{M}_\alpha(t)$. The four graphs of the left-hand column of Figure 4-3 show the time-warped, average traces (from four training subjects) for α every $10°$ from $0°$ to $180°$ for each of the four parameters of the torso stick.[1] A dynamic time warping (DTW) of the 12-dimensional training signals is first performed to align the possibly speed varying curves of a given angle. Because we are using absolute motion, not frame to frame differencing, the amplitude of $\vec{M}_\alpha(t)$ is speed invariant and amenable to scaling by DTW methods. The warped versions are then averaged (shown).

Note how the curves vary slowly as α changes. Since appearance changes slowly with viewpoint, then so does the parameterization. Highly dependent signals such as these can often be well represented by a principal components decomposition, reducing the data to a linear combination of a small number of basis functions. The second column of Figure 4-3 shows all the eigen-functions required to capture 90% of the variance of the warped, averaged torso-stick examples. Notice that the Translation and Rotation parameters need only a single eigen-function and that the Scale parameter requires only two eigen-functions. The right column contains typical reconstructions of the parameter curves using only the eigen-functions shown. Thus, the individual torso-stick traces can be represented using only a small set of eigen coefficients (in fact a singleton in 3 of the 4 parameters). The results are similar for the remaining two sticks. These eigen-functions are stored with the action's BMR so that they may be recalled during the verification procedure (to be discussed). The following section examines recognition using the eigen coefficients within a statistical

[1]The angles $100°$ to $180°$ are generated by reflecting the original image data.

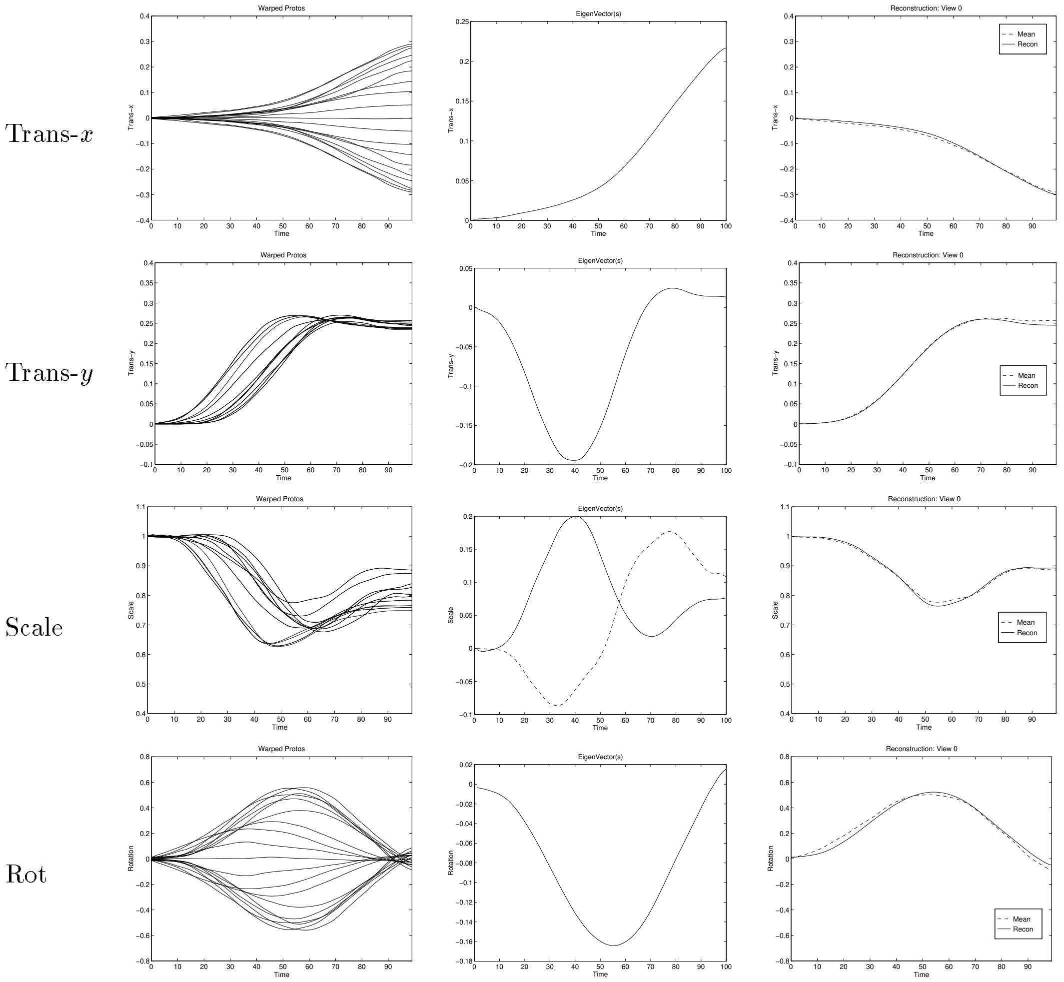


Figure 4-3: Left: Four motion parameters (warped and averaged) of the sitting torso stick for each of training viewing angles; Middle: Eigen-functions capturing 90% of the variance. Right: Typical reconstruction of a motion parameter trace.

framework.

4.1.2 Motion recognition

To determine whether a test motion is consistent with a known movement we need to characterize the variability of the training data. The principal component decomposition discussed above has the effect of reducing the time-based parameter curve to a low-order coefficient vector. As a result, we can capture the variation in the training data by measuring the variation in the eigen coefficients and generating an acceptance region around the coefficients. Figure 4-4 displays the mean value (from training) for each of the five

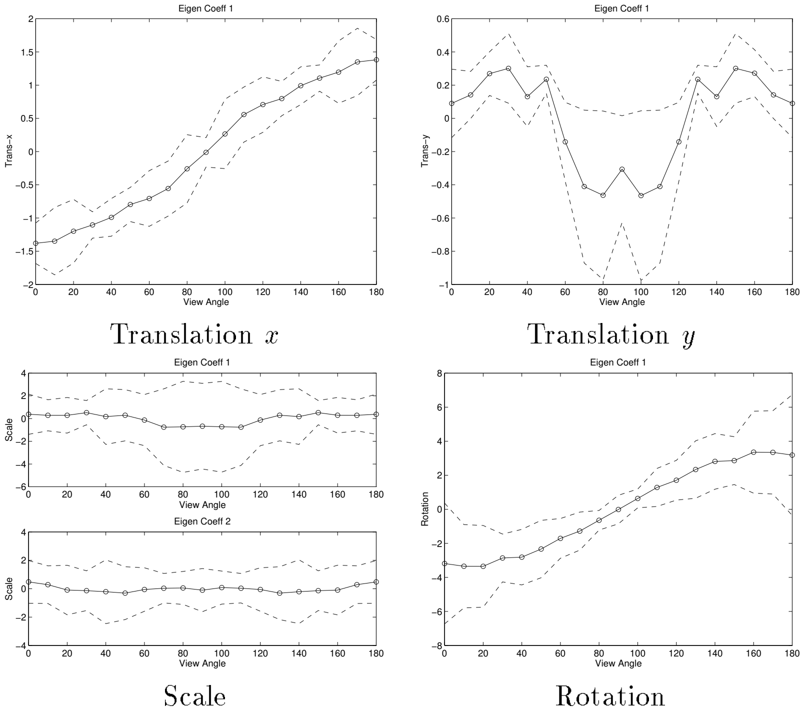


Figure 4-4: Mean and 3σ acceptance region for the torso stick of sitting as a function of the view angle α.

coefficients (two for Scale, and one each for Rotation, Trans-x, Trans-y) for the torso-stick along with a 3σ acceptance envelope. The mean and standard deviation of the coefficients can be considered angle varying vectors $\vec{C}^m(\alpha)$ and $\vec{C}^\sigma(\alpha)$, respectively.

When the BMR hypothesizes a sitting motion at view α_0, the three-stick model, the associated eigen-functions, the mean coefficient vector $\vec{C}^m(\alpha_0)$, and the standard deviation vector $\vec{C}^\sigma(\alpha_0)$ are retrieved. Then the three sticks are placed and tracked, and the necessary parameter trajectories recorded. Next, the input trajectories are jointly dynamically time warped to match the reconstructed trajectories generated by the associated eigen-functions and eigen coefficient mean vector $\vec{C}^m(\alpha_0)$. After warping, the input traces are then projected onto the eigen-function basis set to yield the coefficient vector $\vec{C}^{test}$. Finally, the input motion is accepted as an example of sitting at angle α_0 if every component c_i^{test} of $\vec{C}^{test}$ is within the k-σ envelope of the mean: accept if $\forall_i, \|c_i^{test} - c_i^m\| < k\, c_i^\sigma$, for c_i^m of $\vec{C}^m(\alpha_0)$ and c_i^σ of $\vec{C}^\sigma(\alpha_0)$.

To test this approach we performed the following experiment: We first extracted stick parameters for 3 new people, sitting in different chairs, viewed from the 10 viewing angles.

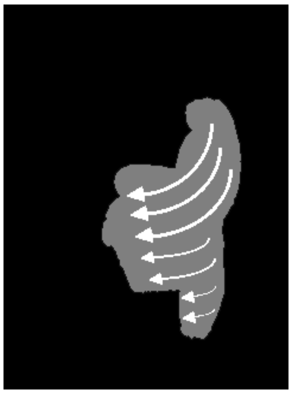


Figure 4-5: Qualitative depiction of the motion field for someone sitting.

We also recorded stick parameters for three aerobic exercises — the closest example to sitting was a simple squat. For a wide range of k, $3.0 \leq k \leq 9.0$, all but one of the sitting examples were accepted, whereas all of the aerobics moves were rejected. This implies that the sitting and aerobic actions were well separated in coefficients-space. Therefore, this approach using motion-based recognition shows promising results, given proper model (stick/patch) tracking.

4.2 Motion Model 2: Static representation of motion

We now examine a second approach to action modeling. Consider the picture shown in Figure 4-5. This image captures the essence of the underlying motion pattern of someone sitting (sweeping down and back) superimposed on the corresponding BMR silhouette. Here, both *where* (BMR silhouette) the motion is happening and also *how* (arrows) the motion is occurring are present in one compact representation. This single image appears to contain the necessary information for determining how a person has moved during the action. Artists have explored this idea of depicting motion using a static image for quite some time [25]. A technique frequently used by artists for representing motion in a single frame is *streaking*, where the motion region in the image is blurred. The result is analogous to a picture taken of an object moving faster than the camera's shutter speed. There are also many interesting photographs in which strobes or moving cameras were used to produce very similar "motion-blurred" effects [9]. In these images the most recent location of motion in the image is shown more prominently than the remaining motion region, much like a viewing the tail of a comet. This "fading" of intensity gives a rough indication of the moving object's path.

In this section, we collapse the temporal motion information into a single image where intensity is a function of recency of motion. The resultant image yields a description similar to the "arrow" picture of Figure 4-5, and it has a visual effect similar to the comet tail with the streaking effect discussed above.

4.2.1 Motion history image (MHI)

To represent how motion is moving, we develop a *motion history image* (MHI). In a MHI, pixel intensity is a function of the motion history at that location, where brighter values correspond to more recent motion. For the results presented here we use a simple replacement and linear decay operator using the binary image difference frame $D'(x, y, t)$:

$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } D'(x, y, t) = 1 \\ \max(0, H_\tau(x, y, t-1) - 1) & \text{otherwise} \end{cases}$$

where the parameter τ represents the maximum length of the time decay. The spatial region of motion remains identical to the BMR. But in accordance to the comet-like description, pixel intensities are now given a linearly ramping value as a function of time, where brighter values highlight the more recent motion locations. It is clear that this replacement operation need not be destructive, and could be made to incorporate previous values by some recursive update.

The image differencing step (to produce $D'(x, y, t)$) in the above procedure can be noisy due to uniform colored areas on the moving object and also due to object/background similarities. One approach to fill the miscellaneous motion gaps is to use image morphology. Unfortunately the dilation and erosion operators used in morphology can be computationally expensive, and we wish to keep real-time performance a concern. Another technique, the one used here, is the use of pyramids (as used in motion estimation [2] and stereo matching [27]) in a hierarchical differencing technique. Though real-time pyramid hardware solutions exist, we used sub-sampled versions of the original image (not low-pass filtered) to generate an approximation to the pyramid. First, the top-level images of the two pyramids (corresponding to two consecutive images) are differenced and thresholded. The resultant image is expanded to the size of the next lower-level images. This lower-level is differenced and

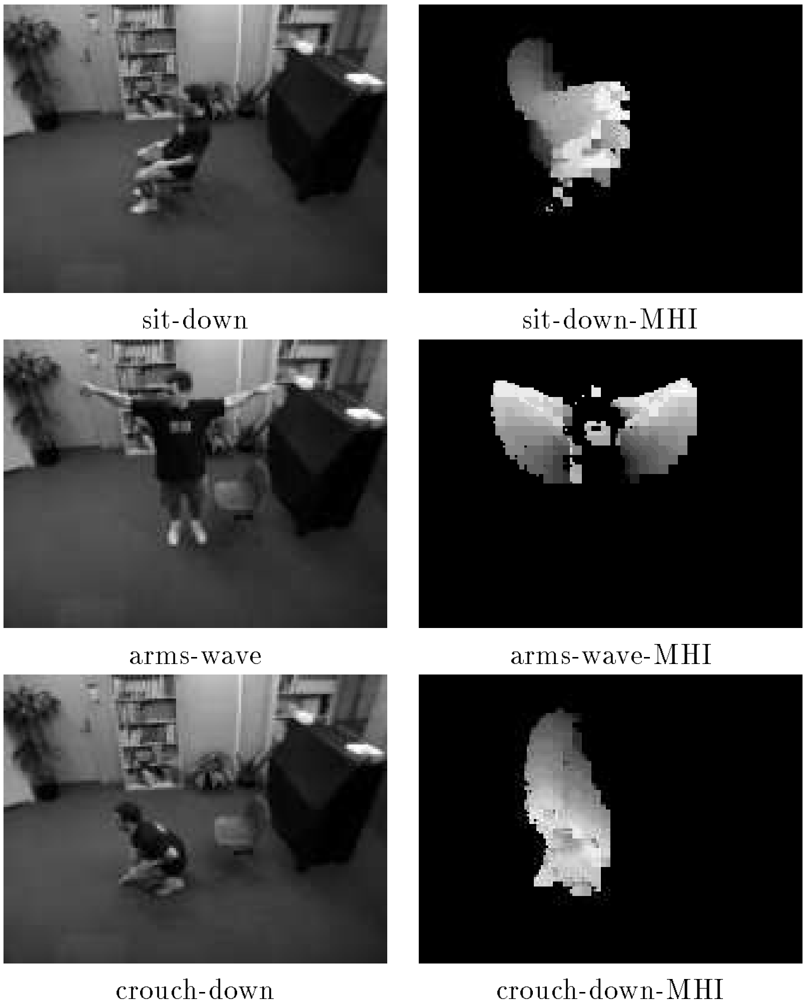


Figure 4-6: Action moves along with their MHIs. The final motion locations in the MHIs appear brighter than the remaining motion regions.

thresholded, and the result is placed into the expanded difference version from the upper-level. This process continues until the bottom-level (the original) images are differenced, thresholded, and incorporated into the expanded difference image. This was an adequate solution to remove the small gaps resulting from image differencing. Examples of MHIs for three actions (sit-down, arms-wave, crouch-down) are presented in Figure 4-6. Notice that the final motion locations appear brighter in the MHIs. Note that the corresponding BMR can be generated by thresholding the MHI above zero. Therefore we use the thresholded version as the BMR for the hypothesis generation, and use the original MHI for testing the targeted actions.

The characterization of the MHI is currently derived from the same moment analysis as the BMR, but now the input image is grayscale instead of binary. Future work will explore a more suitable analysis of this type of image, possibly by modeling the motion history intensity slopes. Since the moment-based features are simple to compute, some may argue that one should bypass the initial hypothesize stage and should match against *all* stored MHI models. Testing this approach with multiple actions showed poorer results than the hypothesize-and-test case (BMR then MHI). Though the BMR shapes may be different, the MHI high-intensity regions (at the end of the action) may be similar. Because the high-energy regions (weighting functions) contribute largely to the moment generation, similar moment sets are created. A benefit of using these grayscale MHIs is that they are sensitive to the direction of motion, unlike the BMRs. Thus, the MHI is especially well-suited for distinguishing between forward and reverse performance of an action (e.g. "sitting down" versus "getting up").

For training the recognition system, we first collect multiple examples of each action from a variety of viewing angles. We then determine the translation- and scale-invariant moment features of the MHIs and compute statistical descriptions (mean, covariance) for each of the actions. Recognition can now be performed given moment features for an unknown movement and the training action statistics.

4.2.2 Real-time action recognition using MHIs

To recognize an input action given a set of moment features and the training statistics, we once again employ the Mahalanobis distance to calculate the relation between the moment description of the input MHI and each of the hypothesized actions. If the Mahalanobis distance to some example is within a given range, then the unknown action can be classified as an instance of that example. A different threshold for each action's acceptance region can be used to accommodate the different statistical distributions.

An important element in this recognition scheme is the temporal segmentation of actions. There is no reason to expect segmentation of the actions in a live video stream. If actions are performed at varying speeds, we need an appropriate time-window length τ for each of the speeds. This length τ determines the extent of the sequence incorporated into the MHI and BMR. During the training phase where we gather example actions, the minimum (τ_{min}) and maximum (τ_{max}) duration that actions may take can be measured. Using these bounds, our current system searches for matches in the video stream using a backwards-looking time

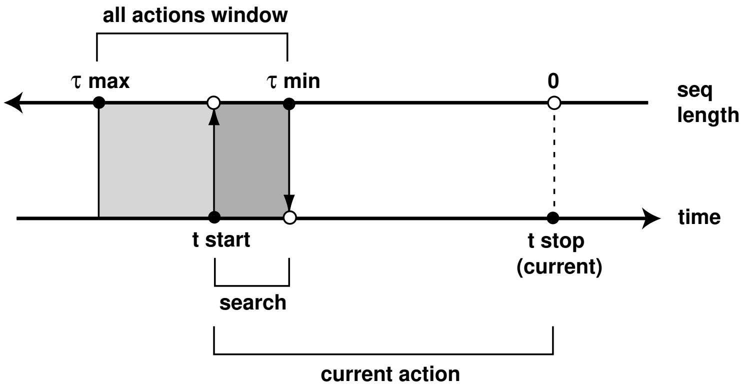


Figure 4-7: Relationship between time and sequence duration for matching actions within a fixed-length search window.

window that resides between τ_{min} and τ_{max} (See Figure 4-7).

Because of the simple nature of the replacement operator used for the MHI, we can construct a highly efficient algorithm for approximating a search over a wide range of τ ($\tau_{min} \leq \tau \leq \tau_{max}$) while retaining only one examination of the image sequence. Basically, truncating values below a given threshold in the MHI and linearly re-normalizing yields a new MHI formed as if using a smaller time window ($\tau < \tau_{max}$) . Therefore, we begin matching on the BMR/MHI set generated with $\tau = \tau_{max}$. Then, we decrease τ towards τ_{min}. As we decrease τ, MHI intensity values below $\tau_{max} - \tau$ are eliminated, and the remaining values are renormalized to be linearly ramping to a maximum of τ (See Figure 4-8). For example, this means that if τ is first set to 30 (most recent), the motion pixel values are assigned a value between 1 and 30. If τ is then decreased to 25 (reducing the time window), we must remove those pixel values which do not belong to this smaller window. These older, decayed motion pixels are in the range 1 to 5. So, truncating pixels values below 5 removes the unwanted data. Then the remaining values are renormalized so they have values between 1 and 25, yielding the linear ramp. Due to real-time constraints, we currently approximate the entire search window by sampling within the search window. Future work will include a more thorough search of the time-window.

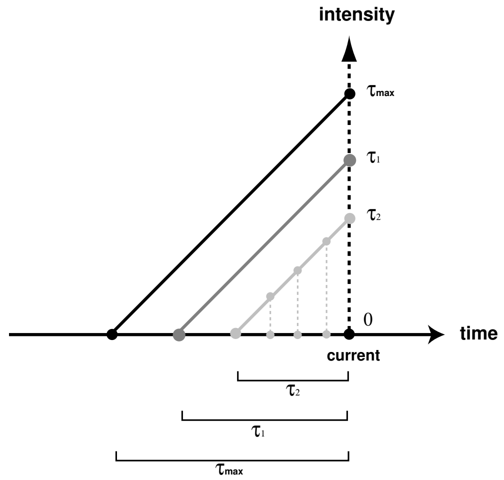


Figure 4-8: For each τ_i window, a truncation and re-normalization of MHI intensity values occurs resulting in a linear ramping of intensity values within the window.

4.3 Extension to multiple views

Many actions can become easily confused using BMRs and MHIs in a single camera scenario. Though it is generally not a cause for concern that their BMRs are similar (in the hypothesize stage, multiple targets are made available), the confusion arises when examining the MHIs. For example, consider sitting and crouching with the person facing the camera. With the destructive motion history operator and a high-positioned, downward-looking camera, the frontal view of these two actions have very similar MHIs. Their brightest highlights are spread nearly all through a similar motion region.

However, the confusion between sitting and crouching does not exist in the side view. Multiple cameras alleviates the difficulty when certain views are easily confused, because there is more information available that can be used to discriminate between actions which appear similar from certain viewing conditions. We incorporated two cameras, set approximately 90° apart (See Figure 4-9), into our a matching approach. The details of the multiple camera recognition method are given in Section 5.2.

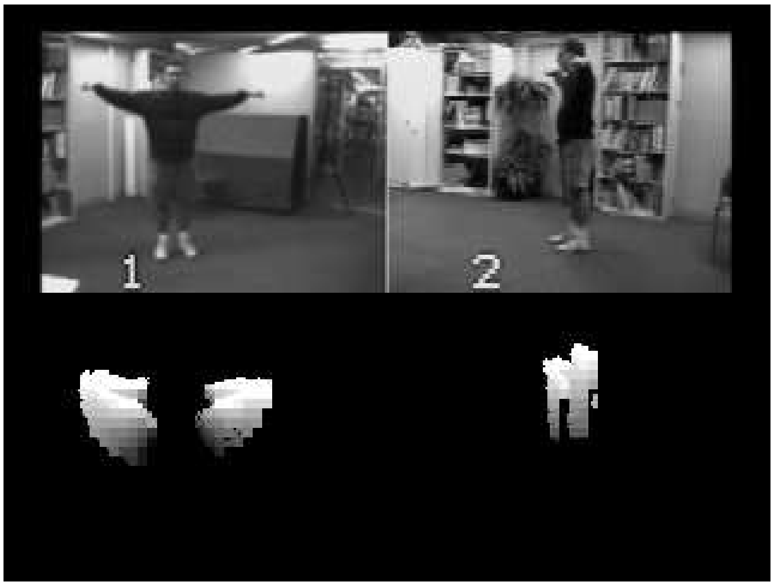


Figure 4-9: Generation of multiple MHIs using multiple cameras. The action is arms-wave (lifting the arms up from the side).

Chapter 5

Results using BMRs and MHIs

5.1 Data

To train the hypothesize-and-test method using BMRs and MHIs we gathered the following data. Four people were asked to perform three actions: sitting, arms-waving, and crouching. Each person performed four repetitions of each action, viewed at 0°, 45°, 90°, 135°, and 180°. Thus, for each of the two cameras, there were 16 examples of every action from 5 view-angle conditions. For both cameras, the action BMRs and MHIs for the training set were calculated, and the moment features were stored in real-time[1]. Then the mean and covariance for the moments are generated for each view of each action. In addition to this training group, one person was used to collect a set of non-action "noise" movements ("walking and moving around within a room") and also a set of reverse actions (e.g. "getting up out of the chair" as opposed to "sitting down"). These false targets were analyzed to manually tune the acceptance thresholds for each action at each view.

5.2 Multiple-camera recognition

In principle, the recognition approach here is designed to permit training using a single camera and recognition with multiple cameras (generally un-calibrated). The fundamental idea is that a single camera acquires the necessary training views (by sampling regions of the view sphere) of the actions to be recognized. Then, multiple cameras are used during

[1] The system runs at approximately 9 Hz (processing 4 time-windows) using two CCD cameras connected to a Silicon Graphics Indy via a Panasonic quad-splitter.

recognition to verify whether the actions seen from each view (of each camera) are consistent with training examples. With no camera calibration, each camera view angle is independent of the others, with no knowledge of the positional relationship between the cameras (i.e. the BMR and MHI for each camera view is treated independently of the other views). If a configuration of cameras is known *a priori* (as in our case, where two cameras are known to be 90° apart), then particular training *view sets* (e.g. 0°\ 90°, 45°\ 135°) would be used, as opposed to a set of random views, for matching with multiple view inputs.

We implemented a three-stage matching procedure to accommodate multiple cameras within the hypothesize-and-test paradigm. The first stage acts as the hypothesis generator and chooses multiple "close" motions to be examined further. The test phase uses stages two and three, where each camera's moment features are separately examined for a tighter match, filtering out the non-actions. If the first stage allows some non-actions to pass, this examination of each camera input separately then permits the salient views to drive the recognition.

The first stage, the hypothesis generator, checks the Mahalanobis distance from the BMR moment sets (one from each view) of an unknown input action to each known action (using a separate mean and covariance for each view). If no camera calibration is known, then every possible view set combination must be examined. But if rough calibration is known, then only a subset of all possible combinations need be searched. To calculate the Mahalanobis thresholds, we initially set the thresholds to be extremely loose (for a wide acceptance region). Using the "noise" BMRs gathered, we reduce the thresholds such that just a few of the noise BMRs are accepted as various actions. This way, the hypothesis construction is generous enough to allow multiple targets to be considered, and not too tight as to reject actions correctly targeted. It is important not to be overly discriminatory during hypothesis generation. In this stage, if there exists a least one *strong* vote for the unknown action to *not* be the proposed example action (e.g. "from this view, the unknown action is definitely not sitting!"), then that hypothesized action is no longer considered for a match. If all the BMRs (one for each view) pass within some threshold, then the second stage is entered.

In the second stage, each BMR moment set (from each camera) is then examined separately using a tighter Mahalanobis threshold. This stage permits the stricter examination of each view independently (e.g. "from this view, the unknown action still resembles sitting").

With non-calibrated cameras, if *any* view supports an example action view, then the action associated with that view is examined for its directional motion properties (MHI) in stage three. If calibrated cameras are used, then a particular view-set passing the first matching stage has its component views examined separately (i.e. if the view-set $0°\backslash\ 90°$ for view-1 \ view-2, passes stage one, then view-1 must pass a stricter match for 0° or view-2 must pass a stricter match for 90°). The thresholds are originally set to be equal to the thresholds set in the first stage. Then each is independently reduced till the noise data fails this second stage. Since the BMR silhouette information is being used, some noise movements (slight movements of entire body) have BMRs similar to crouching. This is because both carve out full-body motion regions. Therefore, correct adjustment of the crouching thresholds remains an issue, and our thresholds were set such that only a few out of several hundred noise movements still passed as crouching. Any BMRs which pass both the stage one and stage two tests are further considered for acceptance in stage three.

As stated earlier, the MHI is particularly well-suited for distinguishing between forward and reverse performance of the actions ("sitting down" versus "getting up"). If any action BMR passes the first two matching stages, then the test action's MHI is examined for a final confirmation of the hypothesized action. A collection of reverse performances of the actions was gathered and used to set the Mahalanobis distance thresholds for the MHI moment sets in stage three. The thresholds here are first set to a large value and then reduced until the reverse moves are no longer accepted. When reducing the thresholds, we found that a few reverse arm-waves were accepted, even when using a small threshold. With some actions, like arms-waving, the motion can be fast and sweep out a small region (as compared to the slower motion of the torso during sitting). This causes some forward and reverse performances to have quite similar MHIs. In case of multiple match successes, we store the actual MHI distances calculated to each target action. Then, after all the targets were examined, the action with the smallest MHI distance (if any) was determined to be the match for the unknown movement.

5.3 Experimental results

To evaluate the hypothesize-and-test method using BMRs and MHIs we conducted the following experiments. To begin, we ran the training data through the three-stage matching

	s0	s45	s90	s135	s180	w0	w45	w90	w135	w180	c0	c45	c90	c135	c180	U
s0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s45	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s90	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0
s135	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0
s180	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0
w0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0
w45	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0
w90	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0
w135	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0
w180	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0
c0	0	0	0	0	0	0	0	0	0	0	14	2	0	0	0	0
c45	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0
c90	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0
c135	0	0	0	0	0	0	0	0	0	0	0	0	0	15	1	0
c180	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0

Table 5.1: Confusion matrix for training data. s:sitdown, w:arms-wave, c:crouch, U:un-classified. Row labels are true nature. Column labels are system classification.

process (See the confusion matrix in Table 5.1). As expected, all actions were correctly classified (e.g. sitting versus crouching versus unknown). It is interesting to note that crouching-0° and crouching-135° chose a nearby view-angle (45° and 180° respectively) for the classification. Given a high correlation between similar view-angle conditions, this confusion is expected.

Next, we had two people perform the three actions. One person (JD) was previously used in generating the training data, and the other person (LC) was not a member of the training group. Each person performed four repetitions of each action, viewed at 0°, 45°, 90°, 135°, and 180°. Their BMRs and MHIs were calculated and their moment sets were classified. Table 5.2 shows the results using these two people for testing. There was no confusion *between* the actions (e.g. labeling "sitting" as "crouching"), and nearby (and symmetric) view-angles were also accepted. But no label could be assigned to some instances of the actions: arm-waving-135°, arm-waving-s180°, crouching-0°, crouching-90°, and crouching-135°). If only the person who also contributed to training (JD) is tested, we notice that the results are much improved (See Table 5.3). In this case, only one unknown match appeared (arm-waving-135°). Thus most of the errors were present in the new test subject (LC), where many movements were un-classifiable (See Table 5.4).

5.4 Discussion

Upon testing our theory, we were initially surprised of the poor results for the subject not included in the training set (LC). We hoped that the four training subjects had enough variability to allow the system to perform well across many individuals. There are a few possible explanations which may account for these results.

	s0	s45	s90	s135	s180	w0	w45	w90	w135	w180	c0	c45	c90	c135	c180	U
s0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s45	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s90	0	0	7	1	0	0	0	0	0	0	0	0	0	0	0	0
s135	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0
s180	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0
w0	0	0	0	0	0	7	0	0	0	1	0	0	0	0	0	0
w45	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0
w90	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0
w135	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	4
w180	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	3
c0	0	0	0	0	0	0	0	0	0	0	3	2	1	0	0	2
c45	0	0	0	0	0	0	0	0	0	0	0	7	0	1	0	0
c90	0	0	0	0	0	0	0	0	0	0	0	0	2	3	1	2
c135	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	1
c180	0	0	0	0	0	0	0	0	0	0	0	0	0	1	7	0

Table 5.2: Confusion matrix for testing data (2 people). s:sitdown, w:arms-wave, c:crouch, U:un-classified. Row labels are true nature. Column labels are system classification.

	s0	s45	s90	s135	s180	w0	w45	w90	w135	w180	c0	c45	c90	c135	c180	U
s0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s45	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s90	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0
s135	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
s180	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
w0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
w45	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
w90	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
w135	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	1
w180	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
c0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0
c45	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0
c90	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0
c135	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0
c180	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0

Table 5.3: Confusion matrix for first test subject (person used in training). s:sitdown, w:arms-wave, c:crouch, U:un-classified. Row labels are true nature. Column labels are system classification.

When gathering the training data, we used 4 people each performing 4 repetitions for a total of 16 examples. Since most people yield fairly consistent repetitions, our approach with only 4 people most likely did not capture enough of the underlying statistics of the actions. It would be more appropriate to have 16 people performing only 1 repetition.

During matching, *all* views are forced to be somewhat "close" to the hypothesized action in the first-level stage of matching. Therefore, any major occlusion present in *any* of the views will cause the termination of further examination for a match (possibly of the correct

	s0	s45	s90	s135	s180	w0	w45	w90	w135	w180	c0	c45	c90	c135	c180	U
s0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s45	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s90	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0
s135	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
s180	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
w0	0	0	0	0	0	3	0	0	0	1	0	0	0	0	0	0
w45	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
w90	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
w135	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	3
w180	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3
c0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	2
c45	0	0	0	0	0	0	0	0	0	0	0	3	0	1	0	0
c90	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2
c135	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	1
c180	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0

Table 5.4: Confusion matrix for second test subject (person not used in training). s:sitdown, w:arms-wave, c:crouch, U:un-classified. Row labels are true nature. Column labels are system classification.

action). This problem must clearly be dealt with if multiple people or objects are to interact within the environment.

It is also noteworthy to mention that grayscale image sequences were used in this implementation. Thus, the image differencing step to build the BMRs and MHIs can be affected (e.g. large motion gaps) by uniform luminance motion regions (as with a moving, solid-colored shirt). Future efforts will include the use of color imagery. Lastly, the data used for the noise and reverse movements were collected using only one person. It is clear that with more noise data and also using multiple people for gathering the reverse data, more precise threshold settings could be determined.

Chapter 6

Summary and extensions

6.1 Summary

From viewing human actions in extremely blurred image sequences, we find that people can readily and easily identify the types of movement. This ability has lead us to develop a motion-based representation of action, instead of feature-based, three-dimensional reconstruction. We showed that two-dimensional motion patterns can be used for the recognition of action and developed two view-based, motion-based approaches for the representation and recognition of action.

A review of relevant research has shown that many approaches to action recognition are divided into two basic paradigms: image-based model construction and motion-based analysis. Many two- and three-dimensional methods for tracking and recognizing human movement define an action as a sequence of statics. There is really no notion of motion, only a series of object or model configurations. We then examined a collection of work on motion recognition based on directional information. This type of approach attempts to characterize the motion itself without reference to the underlying static poses of the body. Selected regions, or patches, of the body are tracked using the motion contained within the regions, and recognition is performed on the dynamic motion changes.

Given a rich vocabulary of activities, an exhaustive search for recognition is generally not feasible. In keeping with the hypothesize-and-test paradigm, we developed a *binary motion region* (BMR) image to act as an index into the known action motions. This index looks at *where* motion is occurring in the image. It is obvious that the silhouetted motion region carved out by someone sitting is *not* similar to someone throwing a ball. Given that a small subset of actions have been targeted by the BMR index, we then examined two methods for representing and recognizing actions based on *how* the motion is occurring. The first method collapses motion parameter trajectories from a tracked stick (or patch) model to a single, low-order vector. Then, for each angle of each move we can define an acceptance region as a probability distribution on that vector. In the second method, we collapse the temporal information of sequential BMRs into a single *motion history image* (MHI), where intensity is a function of recency of motion. The resultant image has an effect similar to motion blurring or streaking. These MHI images have been found to yield reasonable discrimination of multiple actions in a two-camera, real-time system currently in development.

6.2 Extensions

As for present difficulties, there are many extensions required to make the system more adaptable to the complex situations that arise when examining human movements. The first and most obvious extension is to use color, as opposed to grayscale, imagery. The additional chroma values will enhance the image differencing (or motion estimation) used in creating the BMR and MHI generation. We also plan to remove the single person constraint to allow multi-person activities within the environment. Clearly, some form of bounding window(s) will be required. Since collisions or occlusions are frequent in a multi-user environment and are also view-angle specific, multiple cameras can be used in a framework using past history to predict which views to examine. Our current three-stage matching scheme would need to be altered to handle such occlusions.

A more serious difficulty in our motion approach arises when the movement of part of the body is not specified during an action. Consider, for example, throwing a ball. Whether the legs move is not determined by the action itself and induces huge variability in the statistical description of the BMR and MHI. To extend this paradigm to such conditions requires some mechanism to automatically mask away regions of undesired motion.

In keeping with the notion of aspects (or aspect graphs) [23, 21, 10], we desire the generation of view-based *motion* models which can be made to accommodate discrete view regions. In general, the term "aspect recognition" has come to include any recognition scheme that partitions the view sphere into distinct models. The aspect boundaries associated with geometric objects are "hard" divisions. That is, there is no smooth transition between aspects, but a fundamental change in view appearance (e.g. an edge is added or removed). We seek to extend the traditional notion of aspects by partitioning the view-sphere based on motion characteristics, rather than object geometry. Since many motion representations vary smoothly as the viewpoint changes slowly (as shown in this thesis), the aspect boundaries are no longer "hard" divisions — adjacent aspects have a "fuzzy" separation region. Currently, we coarsely sample a $180°$ view arc every $45°$. We hope to define a relationship between the motion models which finely span the view sphere, and to partition that space into coarse aspects based upon some functional criterion. Thus, the motion models accommodating the discrete regions or aspects of the viewing space reduce the magnitude of the number of models necessary for recognition.

We also wish to continue exploring the perception of action in extremely blurred imagery, hoping that such insight will aid in the construction of visually-based perception system for recognizing action. With our current approach, we intend to apply the system to the tasks of detecting particular actions in live video. Eventually, we hope to incorporate our method (and future work) into a real-time interactive environment in which multiple users can participate. We envision systems ranging from a simple "Simon-says" game (where the user mimics the virtual game master and scores, or loses, points) to a virtual aerobics instructor.

Chapter 7

Acknowledgments

There are many people I would like to thank for their help and inspiration. First, I thank my research advisor, Aaron Bobick, for his creative influence. The discussions about this work within our High-Level Vision group (other members include Lee Campbell, Stephen Intille, Claudio Pinhanez, and Andy Wilson) were very helpful. A special thanks goes to those individuals (Aaron Bobick, Lee Campbell, Stephen Intille, Matt Krom, Kate Nguyen, Nuria Oliver, Claudio Pinhanez, Regina Schoonover, Josh Wachman, and Andy Wilson) who offered their time for gathering the action data. I also wish thank Ali Azarbayejani for help with the real-time implementation of the system used for training and testing. In general, I would like to acknowledge all those in the MIT Media Lab's Vision and Modeling group who helped contribute to this work in one form or another. I am also grateful to my thesis readers, Michael Bove and Michael Black, for their time and input on this work. Lastly, I would like to express my appreciation to my friends and family for all their support.

Appendix A

Moment generation

In this appendix, we define two-dimensional moments that are invariant under translation, scale, and orientation. The more complete derivation can be found in [20].

A.1 General moments

The two-dimensional $(p+q)$th order moments of a density distribution function $\rho(x, y)$ (e.g. image intensity) are defined in terms of Riemann integrals as

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q \rho(x, y) dx dy , \tag{A.1}$$

for $p, q = 0, 1, 2, \cdots$.

A.2 Moments invariant to translation

The central moments μ_{pq} are defined as

$$\mu_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \bar{x})^p (y - \bar{y})^q \rho(x, y) d(x - \bar{x}) d(y - \bar{y}) , \tag{A.2}$$

where

$$\begin{aligned} \bar{x} &= m_{10}/m_{00} , \\ \bar{y} &= m_{01}/m_{00} . \end{aligned}$$

It is well known that under the translation of coordinates, the central moments do not change, and are therefore invariants under translation. It is quite easy to express the central moments μ_{pq} in terms of the ordinary moments m_{pq}. For the first four orders, we have

$$
\begin{aligned}
\mu_{00} &= m_{00} \equiv \mu \\
\mu_{10} &= 0 \\
\mu_{01} &= 0 \\
\mu_{20} &= m_{20} - \mu\bar{x}^2 \\
\mu_{11} &= m_{11} - \mu\bar{x}\bar{y} \\
\mu_{02} &= m_{02} - \mu\bar{y}^2 \\
\mu_{30} &= m_{30} - 3m_{20}\bar{x} + 2\mu\bar{x}^3 \\
\mu_{21} &= m_{21} - m_{20}\bar{y} - 2m_{11}\bar{x} + 2\mu\bar{x}^2\bar{y} \\
\mu_{12} &= m_{12} - m_{02}\bar{x} - 2m_{11}\bar{y} + 2\mu\bar{x}\bar{y}^2 \\
\mu_{03} &= m_{03} - 3m_{02}\bar{y} + 2\mu\bar{y}^3
\end{aligned}
$$

A.3 Moments invariant to translation, scale, and orientation

For the second and third order moments, we have the following seven translation, scale, and orientation moment invariants:

$$
\begin{aligned}
\nu_1 &= \mu_{20} + \mu_{02} \\
\nu_2 &= (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2 \\
\nu_3 &= (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \\
\nu_4 &= (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2 \\
\nu_5 &= (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] \\
&\quad +(3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03}) \\
&\quad \cdot[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \\
\nu_6 &= (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]
\end{aligned}
$$

$$
\begin{aligned}
& +4\mu_{11}(\mu_{30}+\mu_{12})(\mu_{21}+\mu_{03}) \\
\nu_7 \quad = \quad & (3\mu_{21}-\mu_{03})(\mu_{30}+\mu_{12})[(\mu_{30}+\mu_{12})^2-3(\mu_{21}+\mu_{03})^2] \\
& -(\mu_{30}-3\mu_{12})(\mu_{21}+\mu_{03})[3(\mu_{30}+\mu_{12})^2-(\mu_{21}+\mu_{03})^2]
\end{aligned}
$$

These moments can used for pattern identification not only independently of position, size, and orientation but also independently of parallel projection.

A.4 Moments invariant to translation and scale

Under the scale transformation for moment invariants we have

$$\mu'_{pq} = \alpha^{p+q+2}\mu_{pq}\,. \tag{A.3}$$

By eliminating α between the zeroth order relation,

$$\mu' = \alpha^2 \mu \tag{A.4}$$

and the remaining ones, we have the following absolute translation and scale moment invariants:

$$\frac{\mu'_{pq}}{(\mu')^{(p+q)/2+1}} = \frac{\mu_{pq}}{\mu^{(p+q)/2+1}}\,, \tag{A.5}$$

for $p+q = 2, 3, \cdots$ and $\mu'_{10} = \mu'_{01} \equiv 0$.

Bibliography

[1] Akita, K. Image sequence analysis of real world human motion. *Pattern Recognition*, 17, 1984.

[2] Bergen, J., P. Anadan, K. Hanna, and R. Hingorami. Hierarchical model-based motion estimation. In *European Conference on Computer Vision*, pages 237–252, 1992.

[3] Black, M. and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motion using local parametric models of image motion. In *ICCV*, 1995.

[4] Campbell, L. and A. Bobick. Recognition of human body motion using phase space constraints. In *ICCV*, 1995.

[5] Cédras, C., and Shah, M. Motion-based recognition: A survey. *Image and Vision Computing*, 13(2):129–155, March 1995.

[6] Cui, Y., D. Swets, and J. Weng. Learning-based hand sign recognition using shoslif-m. In *ICCV*, 1995.

[7] Darrell, T. and A. Pentland. Space-time gestures. In *CVPR*, 1993.

[8] Darrell, T., P. Maes, B. Blumberg, and A. Pentland. A novel environment for situated vision and behavior. In *IEEE Wkshp. for Visual Behaviors (CVPR-94)*, 1994.

[9] Edgerton, H. and J. Killian. *Moments of vision: the stroboscopic revolution in photography.* MIT Press, 1979.

[10] Eggert, D., K. Bowyer, C. Dyer, H. Christensen, and D. Goldgof. The scale space aspect graph. *IEEE Trans. PAMI*, 15(11), 1993.

[11] Essa, I. and A. Pentland. Facial expression recognition using a dynamic model and motion energy. In *ICCV*, June 1995.

[12] Freeman, W. and C. Weissman. Television control by hand gestures. In *Int'l Workshop on Automatic Face- and Gesture-Recognition*, 1995.

[13] Freeman, W., and M. Roth. Orientation histogram for hand gesture recognition. In *Int'l Workshop on Automatic Face- and Gesture-Recognition*, 1995.

[14] Fukumoto, M, Mase, K., and Y. Suenaga. Real-time detection of pointing actions for a glove-free interface. In *IAPR Workshop on Machine Vision Applications*, December 1992.

[15] Gavrila, D. and L. Davis. Tracking of humans in actions: a 3-d model-based approach. In *ARPA Image Understanding Workshop*, Feb 1996.

[16] Goncalves, L., E. DiBernardo, E. Ursella, P. Perona. Monocular tracking of the human arm in 3d. In *ICCV*, June 1995.

[17] Grimson, W. E. *Object Recognition By Computer: The Role of Geometric Constraints.* MIT Press, 1990.

[18] Hoffman, D. and B. Flinchbaugh. The interpretation of biological motion. *Biological Cybernetics*, 45, 1982.

[19] Hogg, D. Model-based vision: a paradigm to see a walking person. *Image and Vision Computing*, 1(1), 1983.

[20] Hu, M. Visual pattern recognition by moment invariants. *IRE Trans. Information Theory*, IT-8(2), 1962.

[21] Ikeuchi, K. and K. S. Hong. Determining linear shape change: Toward automatic generation of object recognition programs. *CVGIP, Image Understanding*, 53(2), 1991.

[22] Ju, S., Black, M., and Y. Yacoob. Cardboard people: a parameterized model of articulated image motion. In *Submitted to the Second International Conference on Automatic Face and Gesture Recognition*, 1996.

[23] Koenderink, and A. van Doorn. The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32, 1979.

[24] Little, J., and J. Boyd. Describing motion for recognition. In *International Symposium on Computer Vision*, pages 235–240, November 1995.

[25] McCloud, S. *Understanding Comics: The invisible art.* Kitchen Sink Press, 1993.

[26] Polana, R. and R. Nelson. Low level recognition of human motion. In *IEEE Workshop on Non-rigid and Articulated Motion*, 1994.

[27] Quam, L.H. Hierarchical warp stereo. *IUW*, 84:137–148, 84.

[28] Rehg, J. and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, 1995.

[29] Rohr, K. Towards model-based recognition of human movements in image sequences. *CVGIP, Image Understanding*, 59(1), 1994.

[30] Shavit, E. and A. Jepson. Motion understanding using phase portraits. In *IJCAI Workshop: Looking at People*, 1995.

[31] Siskind, J. M. Grounding language in perception. In *SPIE*, September 1993.

[32] Ullman, S. Analysis of visual motion by biological and computer systems. *Computer*, August 1981.

[33] Wilson, A. and A. Bobick. Learning visual behavior for gesture analysis. In *Proc. IEEE Int'l. Symp. on Comp. Vis.*, Coral Gables, Florida, November 1995.

[34] Yacoob, Y. and L. Davis. Computing spatio-temporal representations of human faces. In *CVPR*, 1994.

[35] Yamato, J., J. Ohya, and K. Ishii. Recognizing human action in time sequential images using hidden markov models. In *CVPR*, 1992.