Parameterized Modeling and Recognition of Activities

Yaser Yacoob[*]

Michael J. Black[†]

[*] Computer Vision Laboratory, University of Maryland, College Park, MD 20742

[†] Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304

yaser@umiacs.umd.edu, black@parc.xerox.com

Abstract

A framework for modeling and recognition of temporal activities is proposed. The modeling of sets of exemplar activities is achieved by parameterizing their representation in the form of principal components. Recognition of spatio-temporal variants of modeled activities is achieved by parameterizing the search in the space of admissible transformations that the activities can undergo. Experiments on recognition of articulated and deformable object motion from image motion parameters are presented.

Contact

Yaser Yacoob

Computer Vision Laboratory

University of Maryland

College Park, MD 20742

yaser@umiacs.umd.edu

1 Introduction

Activity representation and recognition are central to the interpretation of human movement. There are several issues that affect the development of models of activities and matching of observations to these models,

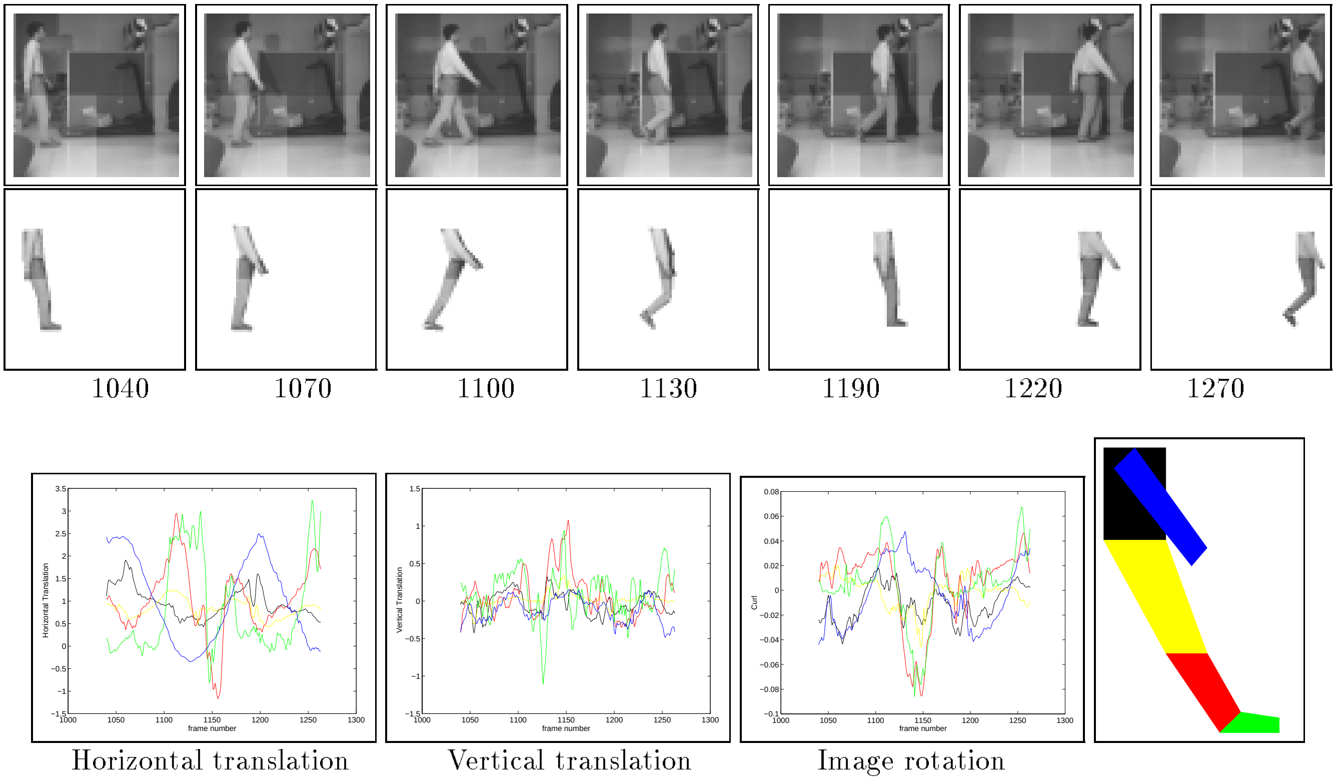
- Repeated performances of the same activity by the same human vary even when all other factors are kept unchanged.
- Similar activities are performed by different individuals in slightly different ways.
- Delineation of onset and ending of an activity can sometimes be challenging.
- Similar activities can be of different temporal durations.
- Different activities may have significantly different temporal durations.

There are also imaging issues that affect the modeling and recognition of activities

- Occlusions and self occlusions of body parts during activity performance.
- The projection of movement trajectories of body parts depend on the observation viewpoint.
- The distance between the camera and the human affect image-based measurements due to the projection of the activity on a 2D plane.

An observed activity can be viewed as a vector of measurements over the temporal axis. The objective of this paper is to develop a method for modeling and recognition of these temporal measurements while accounting for some of the above variances in activity execution.

Consider as an example Figure 1, which shows both selected frames from an image sequence of a person walking in front of a camera and the model-based tracking of five body parts (i.e., arm,



Horizontal translation

Vertical translation

Image rotation

Figure 1: Image sequence of "walking", five part tracking including self-occlusion and three sets of five signals (out of 40) recovered during the activity (torso (black), thigh (yellow), calf (red), foot (green) and arm (blue))

torso, thigh, calf and foot). The figure also shows three (out of eight) motion parameters recovered for each of five body parts (horizontal and vertical translations and rotation in the image plane).

In the remainder of this paper we show that a reduced dimensionality model of activities such as "walking" can be constructed using principal component analysis (PCA, or an eigenspace representation) of example signals ("exemplars"). Recognition of such activities is then posed as matching between the principal component representation of the observed activity ("observation") to these learned models that may be subjected to "activity-preserving" transformations (e.g., change of execution duration, small change in viewpoint, change of performer, etc.).

Figure 2 shows the framework for modeling and recognition of activities. The right side of the figure shows exemplar activities (i.e., instances 1..N of activities) where each instance of an activity has a set of six vectors of temporal measurements. These activities can be modeled using a PCA-based representation as a set of "activity bases" (see lower right part of the figure). The left side of the figure shows an observed activity that is a translated and scaled version of an instance of one of the modeled activities. In this paper we propose an algorithm for recovering the translation, time-scale, and magnitude-scaling of the observed activity given that it is represented in the joint space of activity bases. This algorithm recovers a set of expansion coefficients (i.e., $c_1, ..., c_q$ in Figure 2) that is used in determining the closest matching activity from the exemplars used in learning.

2 Previous Work

Approaches that have been recently employed for modeling and recognizing activities can be divided into data-fitting (e.g., neural networks [15], Dynamic Time Warping (DTW) [7, 8], and regression [12]) feature localization (e.g., scale-space curve analysis [1, 14]) and statistical approaches (e.g., Hidden Markov Models (HMMs) [11, 17]). It is common in these approaches to develop a separate model for each activity, match an observed activity to all models and choose the model that explains it best.

Activity recognition using HMMs was reported in [11, 17] (motion parameters and appearance parameters were used, respectively). In both cases, a set of hidden states was specified a priori and examples were used to estimate the transition probabilities between states. Bobick and Wilson [5] proposed a state-based approach to representing the parameters in an image sequence of gestures.

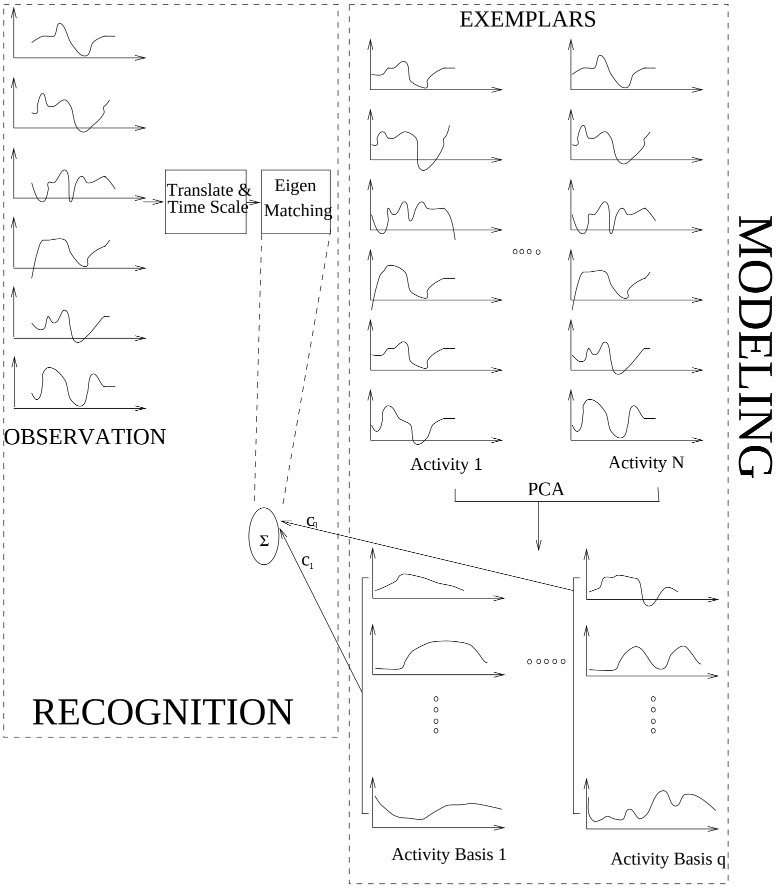


Figure 2: The parameterized modeling and recognition of signals

The states were augmented by a time parameter to preserve temporal ordering. Activity recognition was posed as a search in a space of states representing configurations of gestures using dynamic programming. Some activities have a fine grain continuous structure, not well represented by discrete states. An HMM in which each time instant is represented by a state is more comparable to the continuous representation we develop in this paper.

Recognition of activities subject to "admissible" transformations (e.g., time scaling) enhances the performance of a recognition algorithm since it quantifies the relationship between an instance of an activity and previously encountered instances of that activity. While the above approaches are able to locally handle temporal variability in the data stream of an observed activity, they lack a global detailed model to capture these variabilities. Consequently, it may be difficult with these approaches to explicitly recover and recognize a class of parameterized temporal transformations of an observed activity in respect to a learned model.

Some activities are cyclic thus requiring that several cycles are observed for recognition. Allmen and Dyer [1] proposed a method for detection of cyclic motions from their spatio-temporal curves by tracking high curvature points of the curves. Also, Polana and Nelson [13] proposed an approach to detecting and recognizing activities by low-level spatio-temporal analysis using Fourier transforms. The approach exploits the cyclic nature of some activities to model and recognize them from the motion field measured in image sequences. Seitz and Dyer [16] proposed an approach for determining whether an observed motion is periodic and computing its period. Their approach is based on the observation that the 3D points of an object performing affine-invariant motion are related by an affine transformation in their 2D motion projections.

The approach we propose in this paper is continuous and global (on the time axis) and therefore is an explicit representation of activities. This representation is amenable to matching by global transforms (such as the linear transformation we consider). Also, this global feature allows recognition based on partial or corrupted data (including missing beginning and ending). The most closely related work to the work reported here is that of Bobick and Davis [6] and Ju et al. [9], both proposed using principal component analysis to model parameters computed from activities but did not demonstrate modeling and recognition of activities. Also, Li et al. [10] proposed a PCA-based modeling and recognition approach of whole image sequences of speech.

3 Modeling Activities

Activities will be represented using examples from various activity classes (walking, running etc.). Each example consists of a set of signals. For training, we assume that

- all exemplars are less than or equal to a constant duration
- all examples from a given class are temporally aligned

The $i-$th exemplar from class j is a function from [0...T] on $\mathrm{R^n}$,

$$e^j{}_i(t) : [0..\mathrm{T}] \rightarrow \mathrm{R^n} \tag{1}$$

where n is the number of activity parameters (e.g., translation, rotation etc.) measured at frame t of the image sequence of length T. So, $e^j{}_i(t)$ is a column vector of the n measurements associated with the $j-$th exemplar from activity class i at time t. Let $\bar{e}_i^j$ represent the nT column vector obtained by simply concatenating the $e^j{}_i(t)$ for $t = 0, ..., \mathrm{T}$ into a 1×nT column vector. The set of all j and i of $\bar{e}_i^j$ is used to create the matrix A of dimensions $\mathrm{n\,T} \times \mathrm{k}$ where k being the number of instances of activities $\mathrm{k} < \mathrm{n\,T}$.

Matrix A can be decomposed using Singular Value Decomposition (SVD) as

$$A = U \Sigma V^T \tag{2}$$

where U is an orthogonal matrix of the same size as A representing the principal component directions in the training set. Σ is a diagonal matrix with singular values $\sigma_1, \sigma_2, ..., \sigma_\mathrm{k}$ sorted in decreasing order along the diagonal. The $\mathrm{k} \times \mathrm{k}$ matrix V^T encodes the coefficients to be used in expanding each column of A in terms of principal component directions. It is possible to approximate an instance of activity $\bar{e}$ using the largest q singular values $\sigma_1, \sigma_2, ..., \sigma_q$, so that

$$\bar{e}^{\approx} = \bar{e}^{*} = \sum_{l=1}^{q} c_l U_l \tag{3}$$

where $\bar{e}^{*}$ is the vector approximation, c_l are scalar values that can be computed by taking the dot product of $\bar{e}$ and the column U_l; that is, by projecting the vector $\bar{e}$ onto the subspace spanned by the q basis vectors. The approximation can be viewed as a *parameterization* of the vector $\bar{e}$ in terms of the basis vectors U_l ($l = 1..q$), to be called the *activity basis*, where the parameters are the c_l's.

4 Activity Recognition

Recognition of activities involves matching an observation against the exemplars, where the observation may differ from any of the exemplars due to variations in imaging conditions and performance of activities as discussed earlier. We model variations in performance of an activity by a class of transformation functions $\mathcal{T}$. Most simply, $\mathcal{T}$ might model uniform temporal scaling and time shifting to align observations with exemplars.

Let $\mathbf{D}(t) : [1..\mathrm{T}] \rightarrow \mathrm{R}^{\mathrm{n}}$ be an observed activity and let $[\mathbf{D}]$ denote the nT column vector obtained by first concatenating the n feature values measured at t, for each $\mathbf{D}(t)$ and then concatinating $\mathbf{D}(t)$ for all t. Let also $[\mathbf{D}]_j$ denote the j-th ($j = 1..\mathrm{nT}$) element of the vector $[\mathbf{D}]$. By projecting this vector on the activity basis we recover a vector of coefficients, $\bar{c}$, that approximates the activity as a linear combination of activity bases.

Black and Jepson [3] recently pointed out that projection gives a least squares fit which is not robust. Instead, they employed robust regression to minimize the matching error in an eigenspace of intensity images. Adopting robust regression for recovering the coefficients leads to an error minimization of the form:

$$E(\bar{c}) = \sum_{j=1}^{\mathrm{nT}} \rho(([\mathbf{D}]_j - \sum_{l=1}^{q} c_l U_{l,j}), \sigma) \tag{4}$$

where $\rho(x, \sigma)$ is a robust error norm over x and σ is a scale parameter that controls the influence of outliers. This robustness is effective in coping with random or structured noise. Black and Jepson [3] also parameterized the search to allow an affine transformation of the observation to be used to improve the matching between images and principal images. In our context, a similar transformation allows an observation to be better matched to the exemplars. Let $\mathcal{T}(\bar{a}, t)$ denote a transformation with parameter vector $\bar{a}$ that can be applied to an observation $\mathbf{D}(t)$ as $\mathbf{D}(t + \mathcal{T}(\bar{a}, t))$.

Given an observed activity $\mathbf{D}(t)$, the error minimization of Equation (4) now becomes

$$E(\bar{c}, \bar{a}) = \sum_{j=1}^{\mathrm{nT}} \rho([\mathbf{D}(t + \mathcal{T}(\bar{a}, t))]_j - \sum_{l=1}^{q} c_l U_{l,j}, \sigma) \tag{5}$$

Equation (5) is solved using simultaneous minimization over the coefficient vector $\bar{c}$ and the transformation parameter vector $\bar{a}$. It should be noticed that a more general transformation on $\mathbf{D}(t)$ is possible, specifically $\mathcal{T}(\mathbf{D}(t))$ instead of $\mathbf{D}(t + \mathcal{T}(\bar{a}, t))$. The latter transformation assumes "signal constancy" in terms of the range of values of $\mathbf{D}(t)$ and defines explicitly a "point motion" transformation that is controlled by the model of $\mathcal{T}(\bar{a}, t)$.

The transformed $\mathbf{D}(t + \mathcal{T}(\bar{a}, t))$ can be expanded using a first order Taylor series

$$\mathbf{D}(t + \mathcal{T}(\bar{a}, t)) \approx \mathbf{D}(t) + \mathbf{D}_{\mathrm{t}}(t)\mathcal{T}(\bar{a}, t) \tag{6}$$

where $\mathbf{D}_{\mathrm{t}}$ is the temporal derivative. Equation (5) can be approximated as

$$E(\bar{c},\bar{a}) = \sum_{j=1}^{\mathrm{n\,T}} \rho([\mathbf{D}_t(t)\mathcal{T}(\bar{a},t) + \mathbf{D}(t)]_j - \sum_{l=1}^{q} c_l U_{l,j}, \sigma) \tag{7}$$

Equation (7) can be minimized with respect to $\bar{a}$ and $\bar{c}$ using a gradient descent scheme with a continuation method that gradually lowers σ (see [2]). Initial projection of the observation on the eigenspace provides a set of coefficients $\bar{c}$ that are used to determine an initial estimate of $\bar{a}$ that is used to warp the observation into the eigenspace. The algorithm alternately minimizes the errors of the eigenspace parameterization and the transformation parameterization. Due to the differential term in Equation (7) it is possible to carry out the minimization only over small values of the parameters. To deal with larger transformations a coarse-to-fine strategy can be used to compute the coefficients and transformation parameters at coarse resolution and project their values to finer resolutions similar to what is described in [3]. This coarse-to-fine strategy does not eliminate the need for approximate localization of the curves even at coarse levels.

Upon recovery of the coefficient vector, $\bar{c}$, the normalized distance between the coefficients, c_i, and coefficients of exemplar activities coefficients, m_i, is used to recognize the observed activity. The Euclidean distance, d, between the distance-normalized coefficients is given as

$$d^2 = \sum_{i=1}^{q} (c_i/||\bar{c}|| - m_i/||\bar{m}||)^2 \tag{8}$$

where $\bar{m}$ is vector of expansion coefficients of an exemplar activity. The exemplar activity with the coefficients that score the smallest distance is considered the best match to the observed activity.

5 Experiments

In this section we discuss implementation issues and demonstrate our approach on two different activity domains, articulated and deformable body motions. We show the effectiveness of the proposed approach on large data-sets.

In the first set of experiments, the temporal motion parameters recovered during tracking of a human performing an activity observed from different viewpoints are modeled and then the recognition performance evaluated. The second set focuses on modeling and recognition of four activities as seen from the same viewpoint. Finally, the third set demonstrates the modeling and recognition of speech-reading from visual motion information. Thirteen letters of a single speaker are modeled and recognized using the optical-flow of the mouth motion. In total, several hundred long image sequences of complex activities were used. In these experiments we assume that the objective is recognition of the activity from one cycle (or less) of its performance while ignoring periodicity.

5.1 Modeling and Recognition of Walking

We employ a recently proposed approach for tracking human motion using parameterized optical flow [9]. This approach assumes that an initial segmentation of the body into parts is given and tracks the motion of each part using a chain-like model that exploits the attachments between parts to achieve tracking of body parts in the presence of non-rigid deformations of clothing that cover the parts. The work reported emphasized the low-level tracking component and suggested a possible recognition strategy of the temporal parameters subject to changes of viewpoint and imaging parameters. In this subsection we employ our proposed approach to demonstrate the recognition of activities under varying viewpoints and imaging parameters. We assume that a viewer-centered representation is used for modeling and recognition of several activities. Let $\mathbf{D}(t)$ be the n dimensional signals of an observed activity. A total of five body parts (arm, torso, thigh, calf and foot) were tracked using 8 motion parameters for each part (i.e., n=40). In [9] the observation that the following transformation does not change the the activity $\mathbf{D}(t)$ was made

$$\mathrm{S} * \mathbf{D}(\alpha t + \mathrm{L}) \tag{9}$$

This transformation captures the temporal translation, L, of the curve and the scaling, S, in the magnitude of the signal in addition to the speedup factor α. The magnitude scaling, S, of the signal accounts for different distances between the human and the camera (while the viewing angle is kept constant) and the anthropometric variation across humans. The temporal scaling parameter α is: $\alpha > 1.0$ leads to a linear speed up of the activity and $\alpha < 1.0$ leads to its slow down.

Recognition of activity $\mathbf{D}(t)$ as an instance of a learned activity requires minimizing the error:

$$E(\alpha, \mathrm{L}, \mathrm{S}) = \sum_{j=1}^{\mathrm{nT}} \rho([\mathrm{S} * \mathbf{D}(\alpha t + \mathrm{L})]_j - \sum_{l=1}^{q} c_l U_{l,j}, \sigma) \tag{10}$$

This equation can easily be rewritten and solved as in Equation (7), where

$$\mathcal{T}(\alpha, L, t) = t + (\alpha - 1)t + \mathrm{L} \tag{11}$$

$$E(\bar{c}, \alpha, L, S) = \sum_{j=1}^{\mathrm{nT}} \rho([\mathrm{S} * (\mathbf{D}_t(t)\mathcal{T}(\alpha, L, t) + \mathbf{D}(t))]_j - \sum_{l=1}^{q} c_l U_{l,j}, \sigma) \tag{12}$$

Since the error minimization involves a non-linear term we simplify the computation by observing

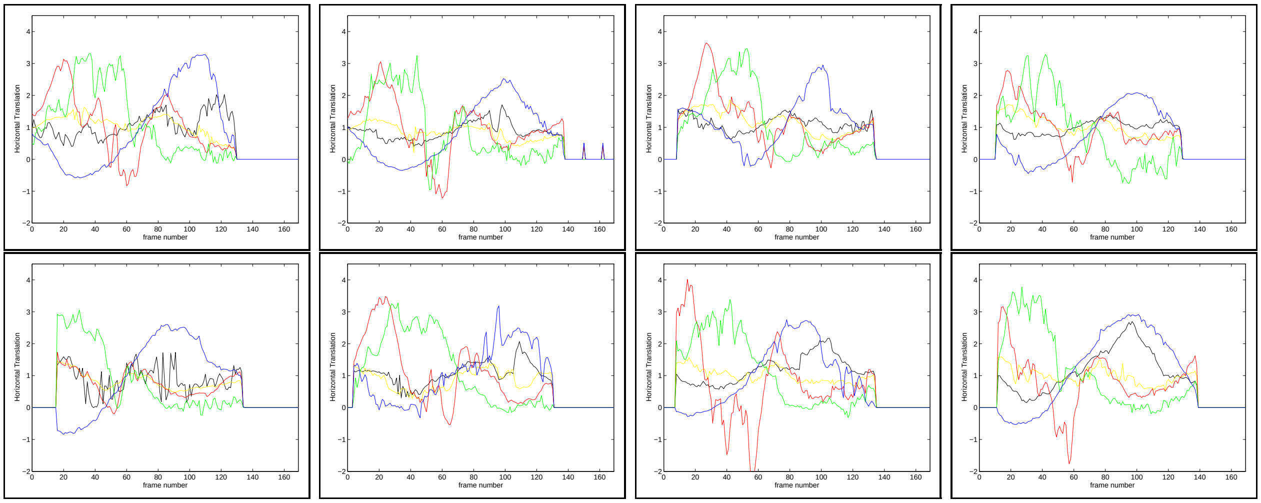


Figure 3: Graphs of eight cycles of "walking" (by different people) showing the horizontal translation parameter of the flow (out of 8) of each of the five patches.

that the multiplication by a constant S can be substituted by dividing the coefficients c_i by S, and therefore in actuality the recovered coefficients are correct up to a scaling factor (i.e., recovering c_i/S. The matching of coefficients is done as in Equation(8). Upon finding the best match the coefficients c_i/S are compared with the matching exemplar coefficients to compute the scaling factor S. Since computing S is overconstrained (q equations with one variable), the mean of S is taken as the scaling factor (i.e., $\mathrm{S} = (\sum_{i=1}^{q}(c_i/m_i))/q$).

The value of S is greater than 1.0 if (a) the activity is viewed at a closer distance than in training (therefore the perception of "larger quantities" is a result of the projection), or, (b) actual faster execution of the activity (which also leads to a temporal scaling for α).

5.2 Synthetic Experiment

In the following experiment we demonstrate the recovery of the parameters of the linear model for a "walking" sequence. We show that unregistered data, with respect to the exemplars, can be aligned using the linear transformation.

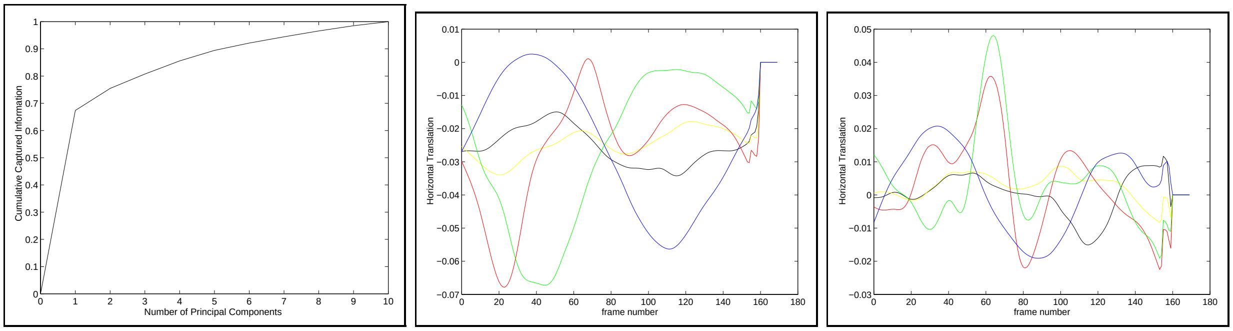


Figure 4: The cumulative information captured as a function of the number of principal components (top) and the first and second principal components (center and right, respectively) for 10 different people walking from a single view for the horizontal translation parameter of the five body parts, (torso (black), thigh (yellow), calf (red), foot (green) and arm (blue)).

Figure 3 shows one temporal parameter (i.e., the horizontal translation) of the five body parts of eight different walkers (out of 10 subjects viewed from the same viewpoint) after the signals have been coarsely registered. The missing cycle parts were filled with "no-activity." Figure 4 shows the first two principal components of one parameter of the walking cycle (however, the forty parameters are modeled in the principal components). Also, the figure shows the ratio of captured information as a function of the number of principal components used in reconstruction (five components are needed to capture 90% of the information while the first component alone captures about 70%). This suggests that a single component can capture walking well if viewed from a single viewpoint.

Figure 5 (bottom) shows five temporal curves of one parameter of a test sequence of a new subject. In this experiment we show the recovery of transformation $\mathcal{T}$ of the "walking." We artificially start the recognition at different frames during the "walking" test sequence (specifically from frame 1015) and recover the translation L and speed α. Notice that the tested activity begins about 35 frames into the "walking" model (Figure 5). A translation of 35 frames will align the tested activity with the model. The graphs in Figure 6 show the recovered translation L and scaling $(\alpha - 1)$ parameters of the "walking" activity as a function of the starting frame. Notice that at

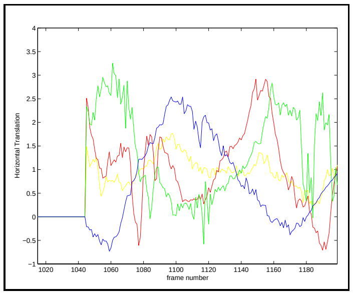


Figure 5: A test sequence used in recognition and evaluation (torso (black), thigh (yellow), calf (red), foot (green) and arm (blue))

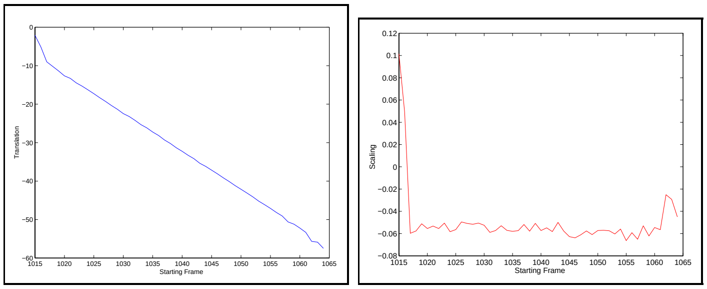


Figure 6: Translation and time scale recovery for the "walking" input curve starting at frame 1015 uptil 1065 (i.e., translating).

frame 1015 a displacement of about 2 frames leftward is needed to align the curve of Figure 5 to the "walking" activity model described in Figure 4. This displacement is increased as the input curve is translated in time. The scaling parameter indicates that the test activity is about 6% faster than the mean "walking" activity. This experiment also shows the effectiveness of the robust norm since it facilitates recognition even when some of the data is inaccurate (e.g., all parameters between frames 1015 and 1045 are zero).

5.3 Multiple-view Modeling and Recognition of Walking

Figure 7 illustrates the experimental set-up for the multi-view walking experiments. The objective here is to demonstrate that a correct classification of the direction of walking of the subjects can be achieved. Since the change in motion trajectories with the change of viewpoint is smooth (see

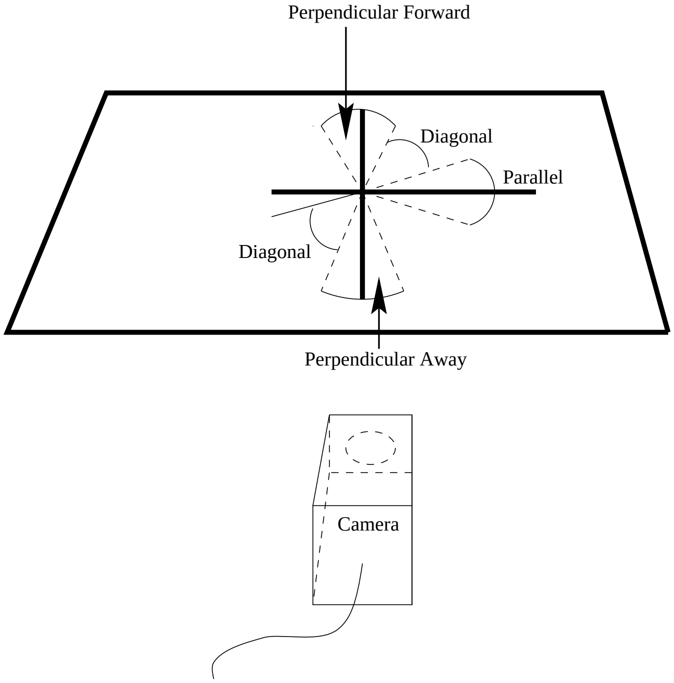


Figure 7: The walking directions in the testing sequences.

[6]) we use four primary directions in the recognition tabulation.

Figure 8 shows the cumulative captured information by the principal components for a single person's walking as viewed from ten different viewing directions (see Figure 7). The angles include walking perpendicular to the camera (towards and away from it). In this case 6 principal components are needed to capture 90% of the information in the motion trajectory of multi-viewpoint observation of walking.

A set of 44 sequences of people walking in different directions were used for testing. The model of multi-view walking was constructed from the walking pattern of one individual while the testing involved eight subjects. The first six activity basis were used. The confusion matrix for the recognition of 44 instances of walking-directions are shown in Table 1. Each column shows the best matches for each sequence. The walkers had different paces and stylistic variations some of which where recovered well by the linear transformation. Also, time shifts were common since only

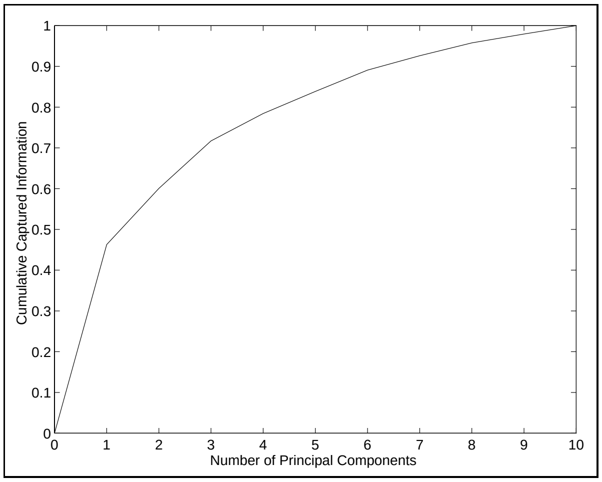


Figure 8: The cumulative information captured as a function of the number of principal components for one person observed walking from 10 different viewing directions.

Walking Direction	Parallel	Diagonal	Perpendicular Away	Perpendicular Forward
Parallel	11	2		
Diagonal	3	14		1
Perpendicular Away			6	
Perpendicular Forward	1	1	1	4
Total	15	17	7	5

Table 1: Confusion matrix for recognition of walking direction

coarse temporal registration was employed prior to recognition. The classification shown in Table 1 was based on the closest distance of the tested data set to a trained viewing direction based on the estimated coefficients.

5.4 Recognition of Four Activities

In this section we illustrate the modeling and recognition of a set of activities that we consider challenging for recognition. We chose four activities that are overall quite close in performance: *walking, marching, line-walking*[1], and *kicking while walking*. Each cycle of these four activities lasts approximately 1.5 seconds.

[1] A form of walking in which the two feet step on a straight line and spatially touch when both are on the ground.

Figures 9-12 show several frames from a performance of each activity by a subject and the tracking of body parts. Also shown are three parameters (for each body part) as measured at each time instant during one cycle. These three parameters are a subset of the eight parameters used in modeling and recognition.

We acquired tens of sequences of subjects performing these four activities as observed from a single view-point. Temporal and stylistic variabilities in the performance of these activities are common. Clothing and lighting variations also affected the accuracy of the recovery of motion measurements from these image sequences. The training sequences were temporally registered so that the beginning of all activities is equal in terms of the perceived configuration of body parts.

Table 2 shows the total number of activities used for both modeling and recognition. The training instances of activities were used to construct the activity basis for the four activities. This activity basis is used in the testing stage on new instances of these activities in which new performers and performances were employed.

Figure 13 (left) shows the percentage of cumulative information captured by the principal components as a function of the number of the principal components for 28 instances of four activities. It also shows how the first three principal components (which capture about 60% -while the fourth principal component captures only 4%) could classify the four activities (see Figure 13 (right), in which the first three expansion coefficients are shown for the 28 activities, the inter-activity variation exceeds the intra-activity variation). In the following recognition experiments, however, we use 15 activity bases to capture most of the information about the activities.

Table 3 shows the confusion matrix for recognition of a set of 66 test activities. These activities were performed by some of the same people who were used for model construction as well as other performers. Variations in performance were accounted for by the linear transformation. Up to 30%

Activity	Number of Training Sequences	Number of Test Sequences
Walking	7	15
Line-Walking	7	28
Marching	7	11
Walking to Kick	7	12

Table 2: List of activities and the number of occurrence of each in training and recognition

Activity	Walking	Line-Walking	Walking to Kick	Marching
Walking	11	3		3
Line-Walking	3	24		1
Walking to Kick			12	
Marching	1	1		7
Total	15	28	12	11

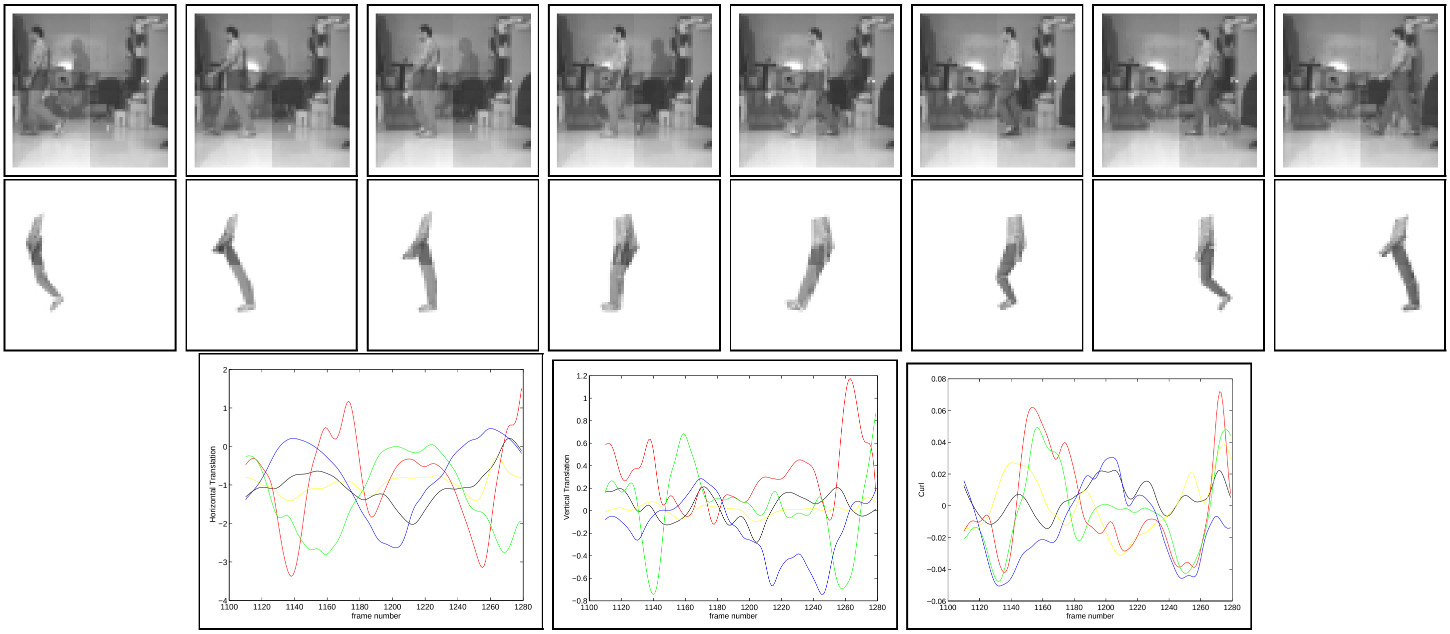
Table 3: Confusion matrix for recognition results

speed-up or slow-down as well as up to 15 frames temporal shift were accounted for by the linear transformation used in the matching.

5.5 Modeling and Recognition of Speech

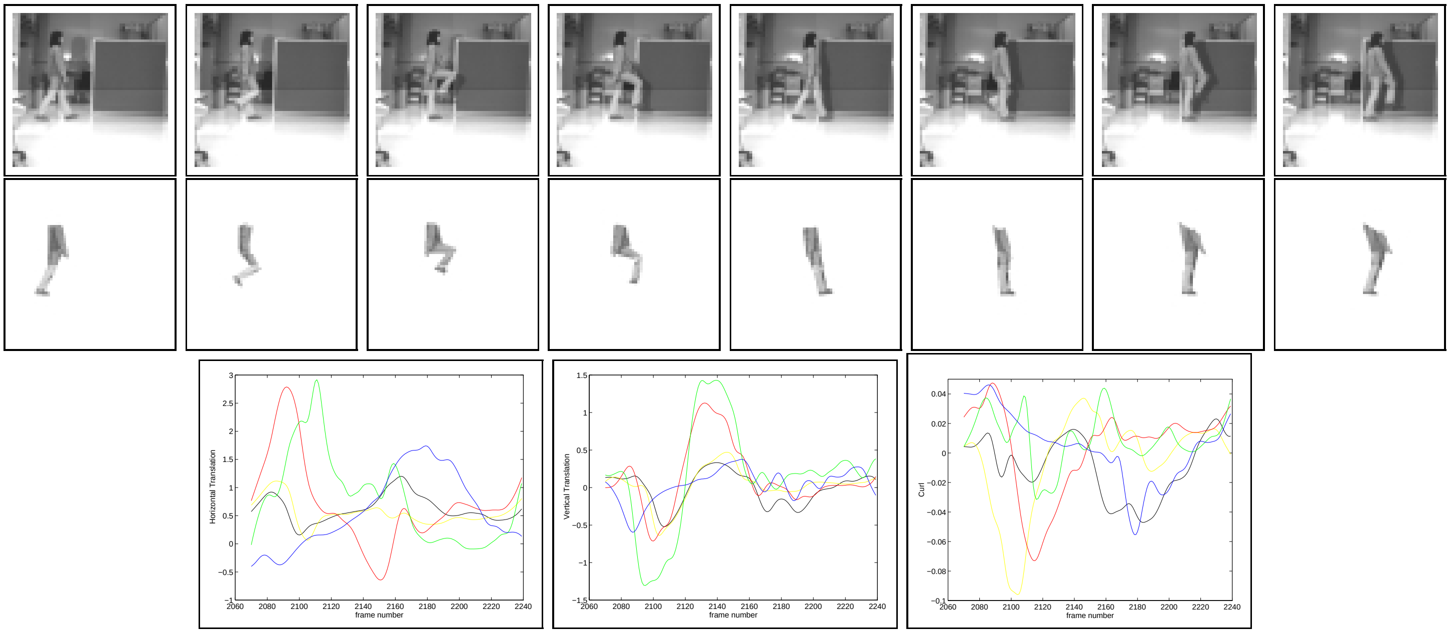
In this section we demonstrate the modeling and recognition of speech from visual information using optical flow measurements computed over long image sequences.

The training set for this experiment consists of 130 image sequences containing a single speaker who utters thirteen letters ten times (Figure 14). The duration of each utterance is 25 frames. We computed the image motion for each sequence in the training set using a robust optical flow algorithm [2]. The robust method is essential as it allows violations of the brightness constancy assumption that occur due to the appearance/disappearance of the teeth, tongue, and mouth cavity. We then randomly chose a subset of 793 flow fields from the training set of 3120 flow fields and derived a low-dimensional representation using principal component analysis (for a detailed description see [4]).



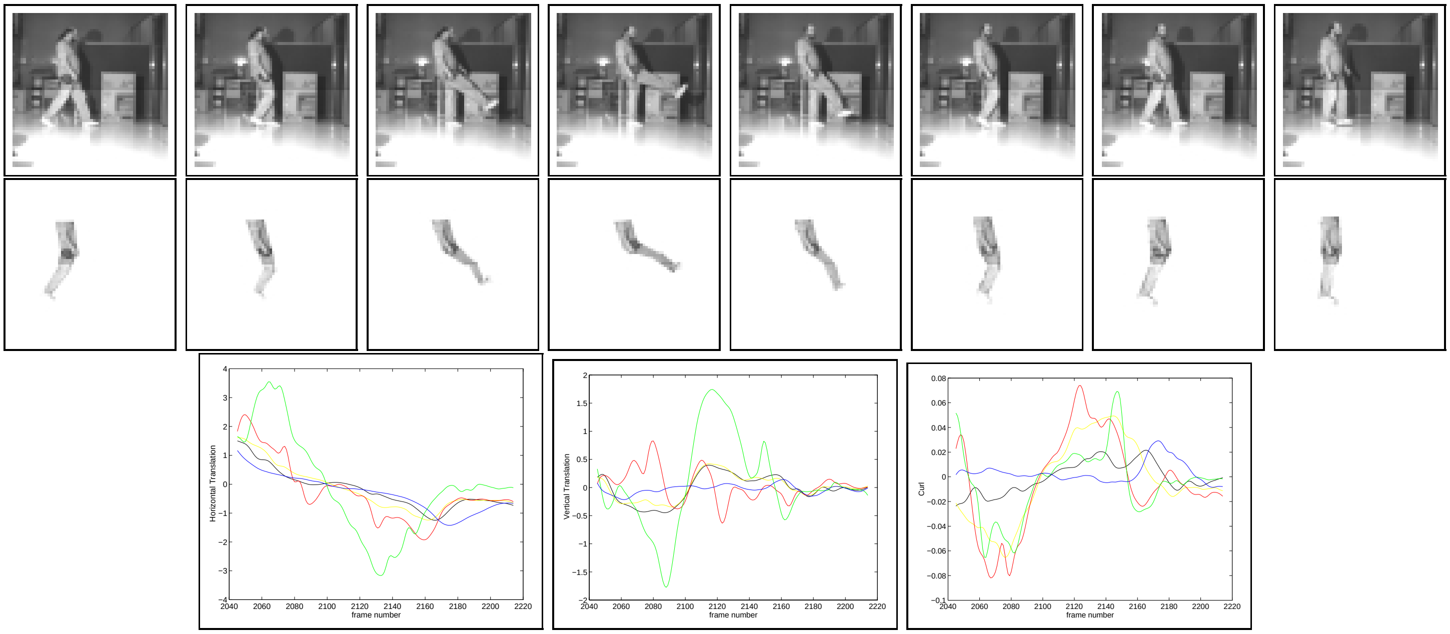
Horizontal translation Vertical translation Rotation in image plane

Figure 9: Image sequence of "walking", five part tracking and three sets of five signals (out of 40) recovered during the activity (torso (black), thigh (yellow), calf (red), foot (green) and arm (blue))



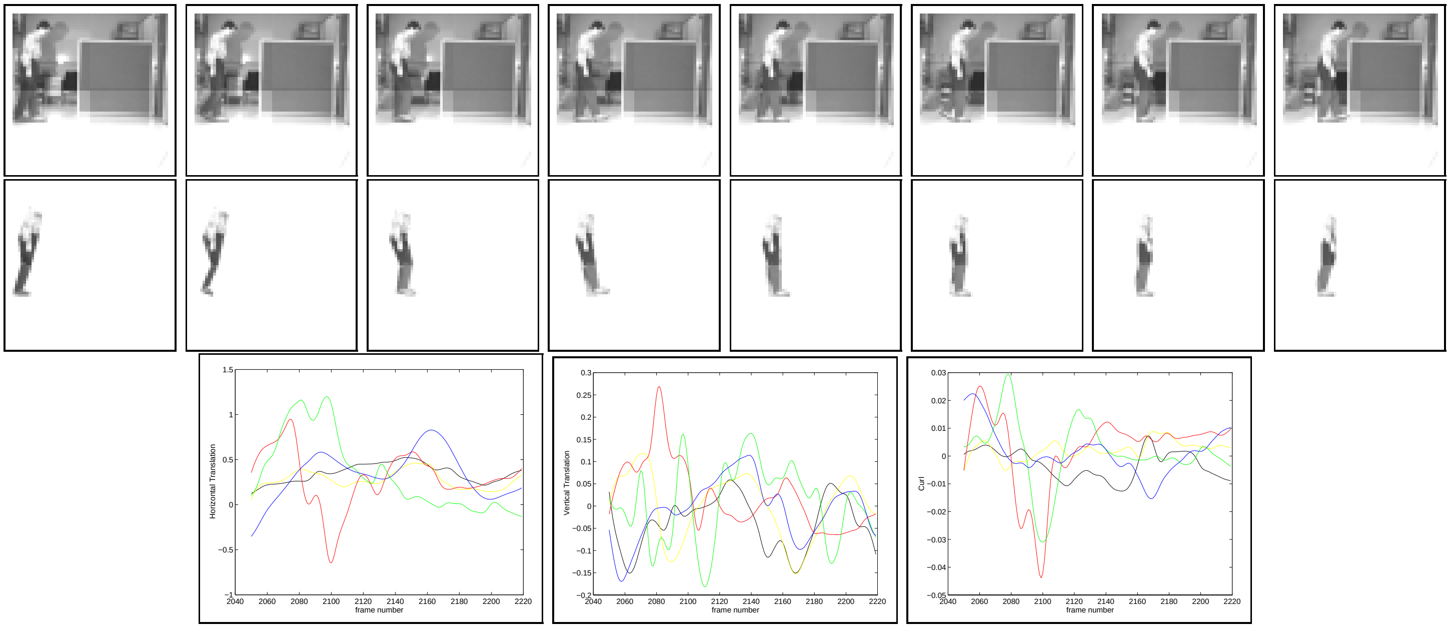
Horizontal translation Vertical translation Rotation in image plane

Figure 10: Image sequence of "marching", five part tracking and three sets of five signals (out of 40) recovered during the activity (torso (black), thigh (yellow), calf (red), foot (green) and arm (blue))



Horizontal translation Vertical translation Rotation in image plane

Figure 11: Image sequence of "kicking", five part tracking and three sets of five signals (out of 40) recovered during the activity (torso (black), thigh (yellow), calf (red), foot (green) and arm (blue))



Horizontal translation Vertical translation Rotation in image plane

Figure 12: Image sequence of "line walking", five part tracking and three sets of five signals (out of 40) recovered during the activity (torso (black), thigh (yellow), calf (red), foot (green) and arm (blue))

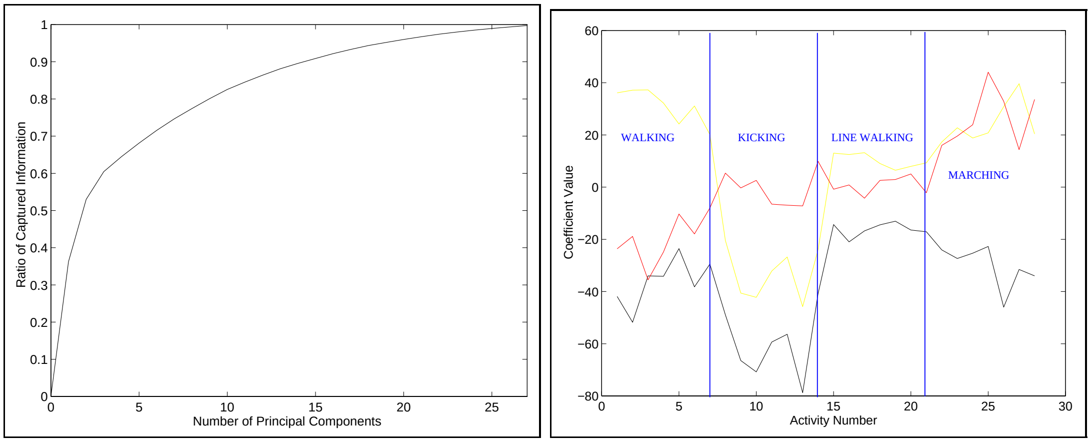


Figure 13: Cumulative information captured by the 28 basis activities (left) and the expansion coefficients of using the first three activity basis for the 28 activities (right) in which classification among activity is clearly visible.

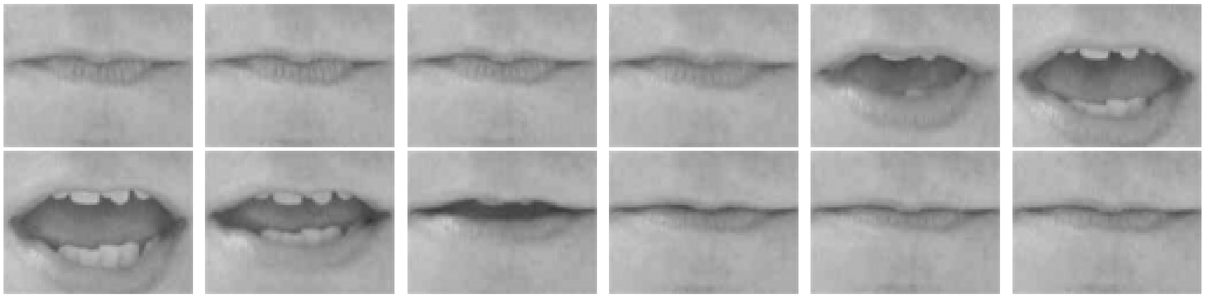


Figure 14: Example frames for one letter in the training set.

Since the image motion of the mouth in our training sequence is constrained, much of the information in the training flow fields is redundant and hence the singular values drop off quickly. For the training data here, the first eight basis flow fields account for over 90% of the information in the training set and are shown in Figure 15.

Image motion is represented as a linear combination of the basis flow templates: $\sum_{i=1}^{8} m_i M_i(\mathbf{x})$ (M_i is a flow template defined over a fixed rectangular region). Using this model, we estimate the motion coefficients m_i as described in [4]. We then use the eight motion coefficients computed between consecutive images to construct a joint temporal model for the letters. We consider each spoken letter to be an activity of 25 frames in duration where eight measurements are computed at

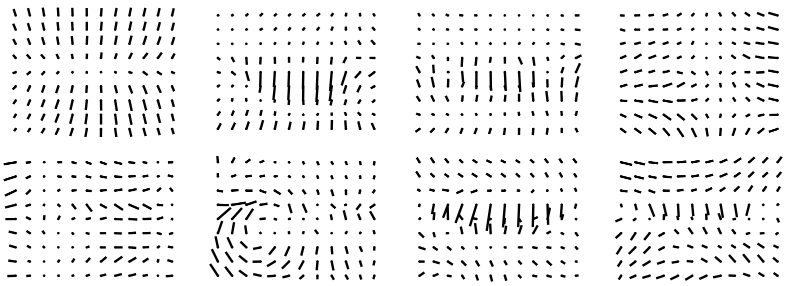


Figure 15: First eight basis flow fields computed by PCA. They account for 90% of the information in the 3120 training flow fields.

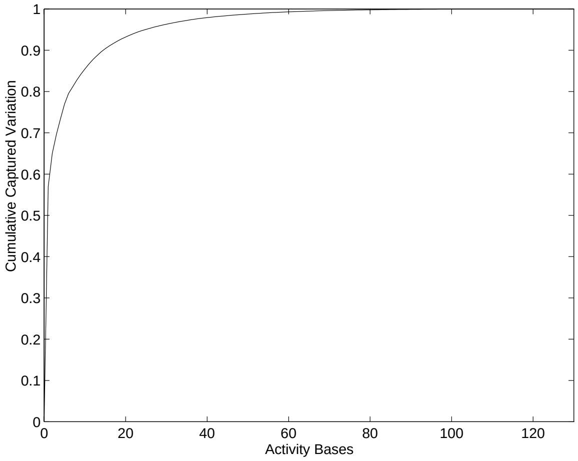
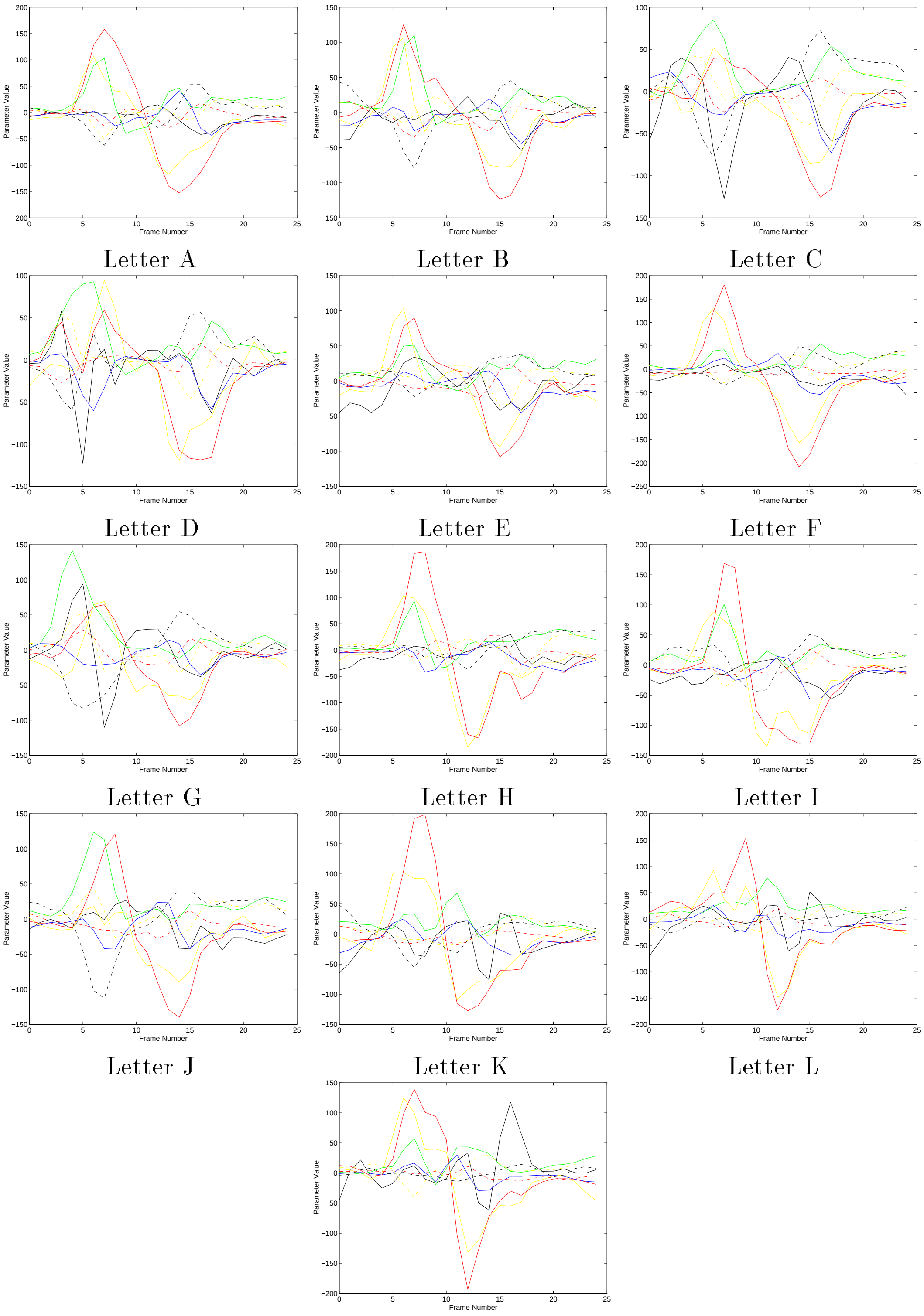


Figure 16: Cumulative variation captured by 130 basis vectors of the 130 sequences

each time instant. The 130 image sequences are used to construct a low-dimensional representation of the 13 letters. These 130 sequences can be represented by a small number of activity-basis as shown in Figure 16. Fifteen activity basis capture 90% of the temporal variation in these sequences.

Figure 17 shows the eight recovered parameters (i.e., the motion-template expansion coefficients) for each letter throughout a single image sequence using a test sequence not in the training set. This figure illustrates the complexity of the modeling and recognition of this large data set.

For the testing of recognition performance, we use 10 new data sets of the same subject repeating the same 13 utterances. A total of 130 sequences were processed. For each two consecutive frames in the test sequences we computed the linear combination of the motion-templates that best describes



Letter A

Letter B

Letter C

Letter D

Letter E

Letter F

Letter G

Letter H

Letter I

Letter J

Letter K

Letter L

Letter M

Figure 17: The eight coefficients of the motion-templates computed for each of 13 letters during a complete utterance

Recognized Letter	A	B	C	D	E	F	G	H	I	J	K	L	M
Letter A	5					1		1	2		2	1	1
Letter B		9		1									
Letter C			6				1			1			
Letter D	1	1	2	5			1						
Letter E					7								
Letter F	2			2		5		1		1			1
Letter G			2	2	1		7			1			
Letter H					1			8					
Letter I	1				1				4		1	3	1
Letter J							1			6			
Letter K						1			4	1	7	1	1
Letter L	1					1						2	
Letter M						2						3	6

Table 4: Confusion matrix for recognition of 130 sequences of 13 letters

the intensity variation (see [4]) and use these parameters in recognition.

The confusion matrix for the test sequences is shown in Table 4. The columns indicate the recognized letter relative to the correct one. Each column sums to 10, the number of each letter's utterances. The confusion matrix indicates that 58.5% correct classification was achieved. When the recognition allowed the correct letter to be ranked second in the matching the success rate increased to 69.3%. Recall that it is well established that visual information is ambiguous for discriminating between certain letters. In this set of experiments we observe some of these confusions. Nevertheless, this experiment shows the effectiveness of the representation we propose for modeling and recognition.

6 Conclusions

In this paper we proposed and tested parametric models for activity modeling and recognition when a large number of temporal parameters are recovered from an image sequence. Principal component analysis and linear transformations were employed to economically represent these activities and effectively recover and recognize instances of learned activities. This approach was demonstarted on large sets of image sequences for recognition of both articulated and deformable motions.

The modeling and recognition algorithm proposed is simple to implement. The principal component analysis determines the proper representation based on the data. Robustness to several sources of variation in performance of activities is an important issue that can be challenging to achieve. The employment of linear transformations in the recognition allowed us to recognize activities even when time scaling and shift were encountered.

The formulation of an activity-preserving transformation can potentially account for a wide range of variations of temporal parameters that result from viewpoint changes and imaging parameters. In this paper we focused on variations of the well understood linear model. The linear transformation, however, is a uniform transformation and therefore is limited to capturing global variations in execution of activities. The formulation we proposed allows future incorporation of non-uniform transformations.

Acknowledgements

We would like to thank Shanon Ju for providing the code for articulated body tracking used in the experiments. Also, we thank Mubarak Shah and Shawn Dettmer for providing the data sets for the speech recognition experiments.

References

[1] M. Allmen and C.R. Dyer. Cyclic Motion Detection Using Spatiotemporal Surfaces and Curves, *ICPR*, 1990, 365–370.

[2] M. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1), 1996, 75–104.

[3] M. J. Black and A. Jepson. EigenTracking: Robust matching and tracking of articulated objects using a view-based representation. *Proc. ECCV* , 1996, 328-342.

[4] M. J. Black, Y. Yacoob, A. Jepson, and D. Fleet. Learning parameterized models of image motion. *Proc. CVPR*, Puerto Rico, June 1997, 561-567.

[5] A. Bobick and A. Wilson. A state-based technique for the summarization and recognition of gesture. *ICCV* , 1995, 382-388.

[6] A. Bobick and J. Davis. An appearance-based representation of action. *ICPR*, 1996, 307-312.

[7] T. Darrell and A. Pentland. Space-time gestures. *Proc. CVPR 93*, 335-340.

[8] D.M. Gavrila and L.S. Davis. Towards 3-D model-based tracking and recognition of human movement: a multi-view approach. *Proc. Workshop on Face and Gesture*, 1995, 272-277.

[9] S. X. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. *Proc. Int. Conference on Face and Gesture*, Vermont, 1996, 561-567.

[10] N. Li, S. Dettmer, and M. Shah. Visually recognizing speech using eigensequences. M. Shah and R. Jain (Eds.), *Motion-Based Recognition*, Klwer Academic Publishing, 1997, 345-371.

[11] C. Morimoto, Y. Yacoob and L.S. Davis. Recognition of head gestures using Hidden Markov Models *International Conference on Pattern Recognition*, Vienna, Austria, August 1996, 461-465.

[12] S.A. Niyogi and E.H. Adelson. "Analyzing and recognizing walking figures in XYT." *CVPR*, 1994, 469-474.

[13] Polana and R. Nelson, Detecting Activities. *IEEE CVPR*, 1993, 2-7.

[14] K. Rangarajan, W. Allen and M. Shah. Matching motion trajectories using scale-space. *Pattern recognition*, 26(4), 1993, 595-610.

[15] M. Rosenblum, Y. Yacoob and L.S. Davis. Human Expression Recognition from Motion Using a Radial Basis Function Network Architecture, *IEEE Transactions on Neural Networks*, Vol. 7, No. 5, 1996, 1121-1138.

[16] S.M. Seitz and C.R. Dyer. Affine Invariant Detection of Periodic Motion, *IEEE CVPR*, 1994, 970-975.

[17] T. Starner and A. Pentland. Visual Recognition of American Sign Language Using Hidden Markov Models. In *International Workshop on Automatic Face and Gesture Recognition,* 1995, 189-194.