Computer Vision and Image Understanding
Vol. 73, No. 2, February, pp. 232–247, 1999
Article ID cviu.1998.0726, available online at http://www.idealibrary.com on IDEAL®

Parameterized Modeling and Recognition of Activities

Yaser Yacoob

Computer Vision Laboratory, University of Maryland, College Park, Maryland 20742
E-mail: yaser@umiacs.umd.edu

and

Michael J. Black

Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, California 94304
E-mail: black@parc.xerox.com

Received October 13, 1997; accepted July 27, 1998

In this paper we consider a class of human activities—atomic activities—which can be represented as a set of measurements over a finite temporal window (e.g., the motion of human body parts during a walking cycle) and which has a relatively small space of variations in performance. A new approach for modeling and recognition of atomic activities that employs principal component analysis and analytical global transformations is proposed. The modeling of sets of exemplar instances of activities that are similar in duration and involve similar body part motions is achieved by parameterizing their representation using principal component analysis. The recognition of variants of modeled activities is achieved by searching the space of admissible parameterized transformations that these activities can undergo. This formulation iteratively refines the recognition of the class to which the observed activity belongs and the transformation parameters that relate it to the model in its class. We provide several experiments on recognition of articulated and deformable human motions from image motion parameters.

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.
- In the modeling stage, defining the activity from onset to offset can sometimes be challenging. While in the recognition stage the onset and ending of an activity must be determined in conjunction with activity identification.
- 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 human 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 modeled using vectors of measurements at discrete time instants that capture the motion of body parts. 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 Fig. 1 (see Appendix for computation details), 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, torso, thigh, calf, and foot [11]). The figure also shows two motion parameters recovered for each of the five body parts (horizontal translation and rotation in the image plane).

In this paper we address a class of activities that we label "atomic" activities. These are defined to be human movements that satisfy the following:

- The movements are structurally similar over the range of performers. For example, a cycle of walking is an atomic activity since its execution steps are quite similar among people and its speed varies within known ranges defined by physical constraints. In contrast, a "jump" movement does not have a single structure since people may jump on their left, right, or both legs, or vertically only or horizontally and vertically.
- The movements are mapped onto a finite temporal window. For example, a cycle of walking is an atomic activity since it can be mapped onto a finite temporal window that is bounded by a maximal walking speed. Periodic movements are not atomic

1077-3142/99 $30.00

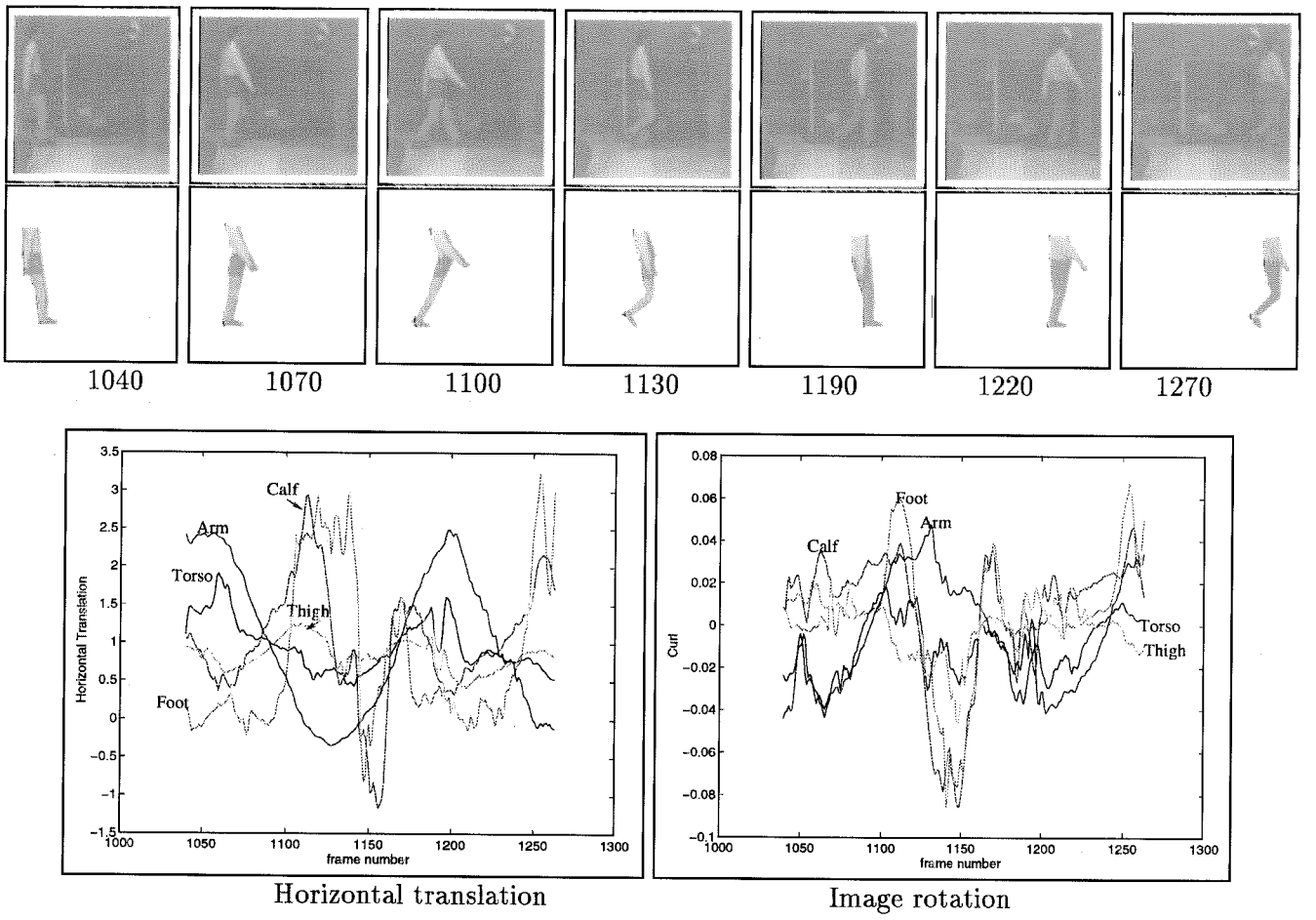

FIG. 1. Frames from an image sequence of walking (top row), five parts tracking of the visible human body parts (arm, roso, thigh, calf, and foot, second row), and two sets of five signals (out of 40), horizontal translation and image rotation that are recovered during the activity (torso, thigh, calf, foot, and arm).

since they can be composed of many cycles and thus are not time bounded.

Modeling and recognition of atomic activities are most challenging when the activities have both structural similarities and similar temporal durations (e.g., "walking" versus "marching" cycles). In this case, a comprehensive modeling and recognition strategy is needed. In this paper we focus on these activities and provide two sets of atomic activities that humans regularly perform; the first involves articulated movements such as "walking," "kicking," and "marching," while the second involves deformable mouth motions during single letter utterances.

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 illustrates the framework for modeling and recognition of activities. The right side of the figure shows exemplar activities (i.e., instances $1..k$ of M different atomic activities, $k > M$) where each instance of an activity has a set of six signals of temporal measurements. These activities can be modeled using a PCA-based representation as a set of q "activity bases," $q \ll k$ (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., $c_1, \ldots, c_q$ in Fig. 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.,

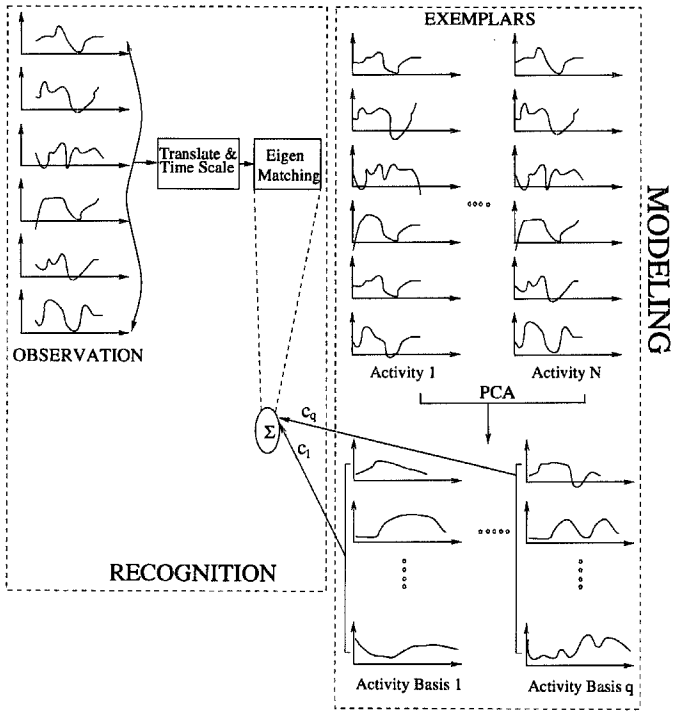

FIG. 2. The parameterized modeling and recognition of measurement signals of activities.

neural networks [17], dynamic time warping (DTW) [9, 10], regression [14]), feature localization (e.g., scale-space curve analysis [1, 16]), and statistical approaches (e.g., hidden Markov models (HMMs) [8, 13, 19]). 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 [8, 13, 19] based on motion and appearance parameters. In these cases, a set of hidden states was specified a priori and examples were used to estimate the transition probabilities between states. Bobick and Wilson [6] proposed a state-based approach to representing the parameters in an image sequence of gestures. 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 sparse discrete states. An HMM in which each time instant is represented by a state is more comparable to the 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 with respect to a learned model.

Some activities are cyclic, thus requiring that several cycles be 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 [15] 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 image motion (normal flow) measured in image sequences. Seitz and Dyer [18] 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 time-contiguous and global; therefore, it 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 onset or offset data). The most closely related work to the work reported here is that of Bobick and Davis [7] and Ju *et al.* [11]; 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.* [12] proposed a PCA-based modeling and recognition approach that exploited entire image sequences of people utterances.

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 jth exemplar from class i is a function from $[0 \ldots T]$ on R^n,

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

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, $[e_i^j(t)]$ is a column vector of the n measurements associated with the jth exemplar from activity class i at time t. Let $\bar{e}_i^j = [e_i^j]_{t=0}^T$ represent the nT column vector obtained by simply concatenating the $e_i^j(t)$ for $t = 0, \ldots, T$ into a $1 \times nT$ column vector. The set of all j and i of $\bar{e}_i^j$ is used to create the matrix A of dimensions $nT \times k$, where k is the number of instances of all M activities, $k < nT$.

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 $\sigma_1, \sigma_2, \ldots, \sigma_k$, sorted in decreasing order along the diagonal. The $k \times k$ 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 $\sigma_1, \sigma_2, \ldots, \sigma_q$, so that

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

where $\bar{e}^*$ is the vector approximation and 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 ($l = 1 \ldots q$), to be called the *activity bases*, 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 $\mathbf{D}(t) : [1 \ldots T] \rightarrow R^n$ be an observed activity and let $[\mathbf{D}]$ denote then nT column vector obtained by first concatenating the n feature values measured at t, for each $\mathbf{D}(t)$ and then concatenating $\mathbf{D}(t)$ for all t. Let also $[\mathbf{D}]_j$ denote the jth ($j = 1 \ldots nT$) element of the vector $[\mathbf{D}]$. By projecting this vector on the activity basis we recover a vector of coefficients, $\vec{c}$, 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

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

where $\rho(x, \sigma)$ is a robust error norm over x and σ is a scale parameter that controls the influence of outliers. In the experiments in this paper we use

$$\rho(x, \sigma) = \frac{x^2}{\sigma^2 + x^2}.$$

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 $\mathcal{T}(\vec{a}, t)$ denote a transformation with parameter vector $\vec{a}$ that can be applied to an observation $\mathbf{D}(t)$ as $\mathbf{D}(t + \mathcal{T}(\vec{a}, t))$.

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

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

Equation (5) is solved using simultaneous minimization over the coefficient vector $\vec{c}$ and the transformation parameter vector $\vec{a}$. It should be noticed that a more general transformation on $\mathbf{D}(t)$ is possible, specifically $\mathcal{T}(\mathbf{D}(t))$ instead of $\mathbf{D}(t + \mathcal{T}(\vec{a}, t))$. We chose the latter transformation since it imposes "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 $\mathcal{T}(\vec{a}, t)$.

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

$$\mathbf{D}(t + \mathcal{T}(\vec{a}, t)) \approx \mathbf{D}(t) + \mathbf{D}_t(t)\mathcal{T}(\vec{a}, t), \tag{6}$$

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

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

Equation (7) can be minimized with respect to $\vec{a}$ and $\vec{c}$ 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 $\vec{c}$ that are used to determine an initial estimate of $\vec{a}$ 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 Eq. (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, $\vec{c}$, 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

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

where $\vec{m}$ 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 [11] (see Appendix). 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 nonrigid 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

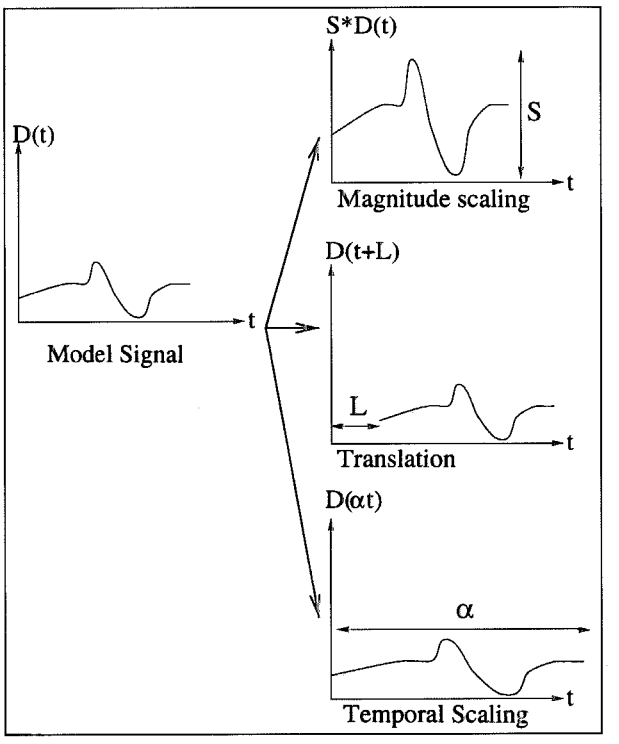

FIG. 3. The effect of each parameter of the transformation on a signal $D(t)$. The magnitude scaling (top right), temporal translation (center right), and temporal scaling (bottom right) of $D(t)$.

using eight motion parameters for each part (i.e., $n = 40$). In [11] it was suggested that the following transformation does not change the activity $\mathbf{D}(t)$:

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

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 $\alpha > 1.0$ leads to a linear speed up of the activity and $\alpha < 1.0$ leads to its slow down. Figure 3 describes the effect of each parameter on a single signal.

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

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

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

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

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

Since the error minimization involves a nonlinear term we simplify the computation by observing 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 Eq. (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., $S = (\sum_{i=1}^{q}(c_i/m_i))/q$).

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) there is an 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 shows the first two principal components of one parameter of the walking cycle, horizontal translation a_0 (however, the 40 parameters are modeled in the principal components) for sample walking cycles from 10 subjects viewed from the same viewpoint. Also, the figure shows the ratio of captured information, $\sum_{l=1}^{q} \lambda_l^2 / \sum_{l=1}^{k} \lambda_l^2$ (λ_l is the lth largest eigenvalue), 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

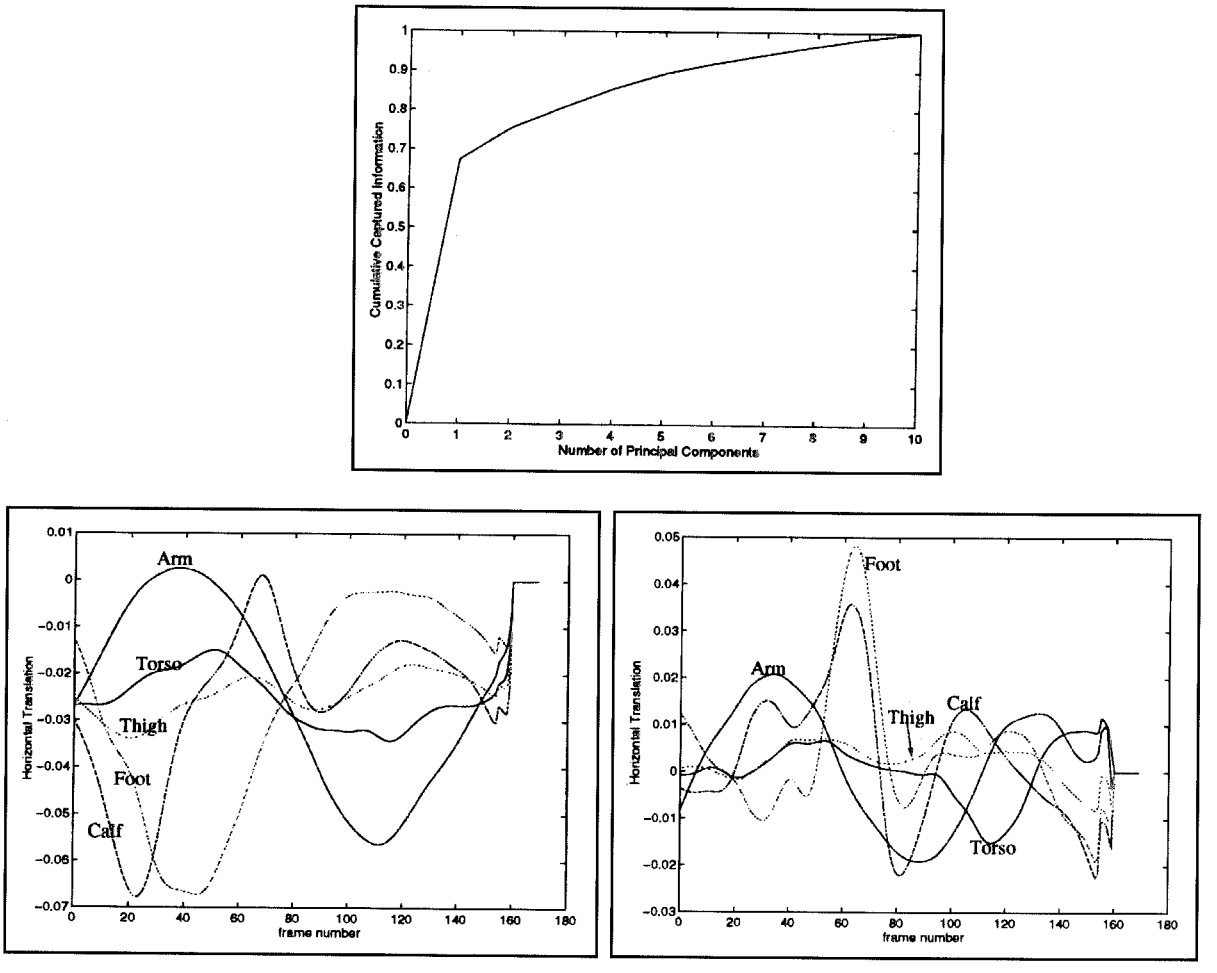

FIG. 4. The cumulative information captured as a function of the number of principal components (top) and the first and second principal components (left and right, respectively) for 10 different people walking from a single view for the horizontal translation parameter of the five body parts, (torso,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

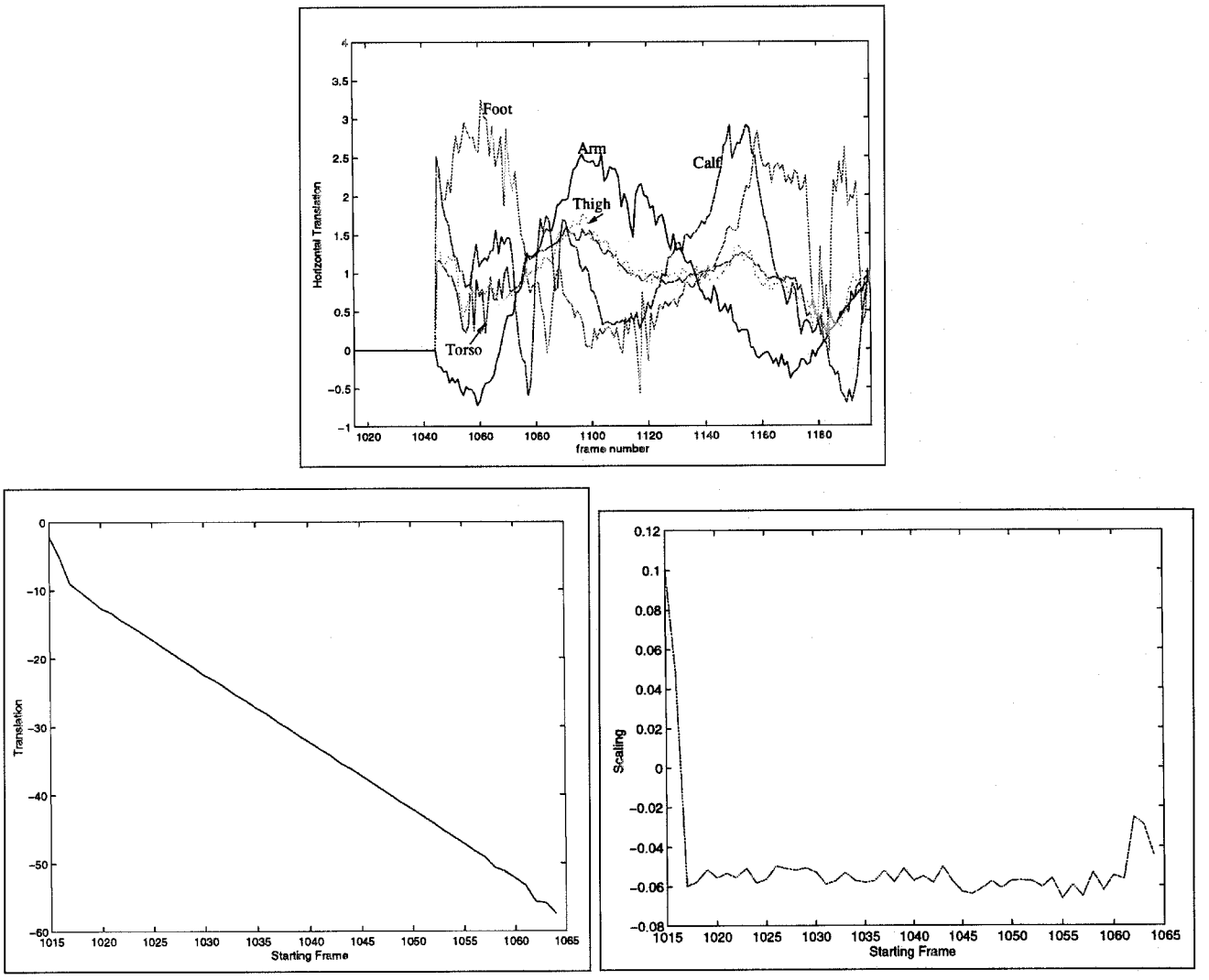

FIG. 5. The horizontal motion parameter a_0 of the five tracked body parts of a test sequence used in recognition and evaluation (left graph) and temporal translation and time scale recovery for the "walking" input curve starting at frame 1015 until frame 1065 (i.e., translated).

70%). This suggest that a single component can capture walking well if viewed from a single viewpoint.

Figure 5 shows five temporal curves of one parameter a_0 of a test sequence of a new subject. In this experiment we show the online recovery of transformation $\mathcal{T}$ for 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 (Fig. 5). A translation of 35 frames will align the tested activity with the model. The other two graphs in Fig. 5 show the recovered translation L and scaling $(\alpha - 1)$ parameters of the walking activity as a function of the starting frame. Notice that at frame 1015 a displacement of about two frames leftward is needed to align the curve of Fig. 5 to the walking activity model described in Fig. 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 due to noise or because it is missing (e.g., all parameters between frames 1015 and 1045 are zero).

Figure 7 shows the cumulative captured information by the principal components for a single person's walking as viewed from 10 different viewing directions (see Fig. 6). The angles include walking perpendicular to the camera (toward and away from it). In this case six principal components are needed to capture 90% of the information in the motion trajectory of a multi-viewpoint observation of walking. Figure 8 shows frames from test sequences for four walking directions.

A set of 44 sequences of people walking in different directions was used for testing. The model of multiview walking was constructed from the walking pattern of one individual while the testing involved eight subjects. The first six activity bases were

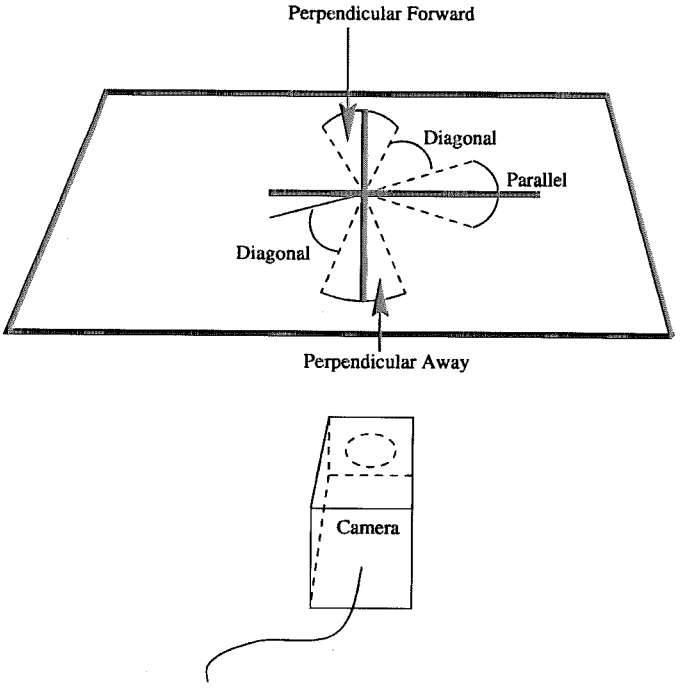

FIG. 6. The experimental setup in the multiview walking sequences.

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 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.

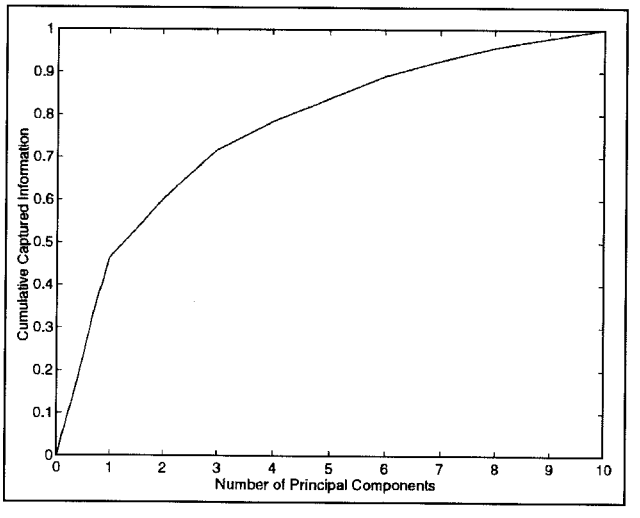

FIG. 7. The cumulative information captured as a function of the number of principal components for one person observed walking from 10 different viewing directions.

TABLE 1
Confusion Matrix for Recognition of Walking Direction

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

5.3. 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 s.

Figure 9 shows several frames from a performance of each activity by a subject and the tracking of body parts. We acquired tens of sequences of subjects performing these four activities as observed from a single viewpoint. 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 10 (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 Fig. 10 (right), in which the first three expansion coefficients are shown for the 28 activities; the interactivity variation exceeds the intra-activity variation). Recall that the coefficients of the training examples are computed by projecting each activity in the training set onto each one of the basis activities using scalar multiplication. The labels point to the four types of activities used in the training set. 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

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

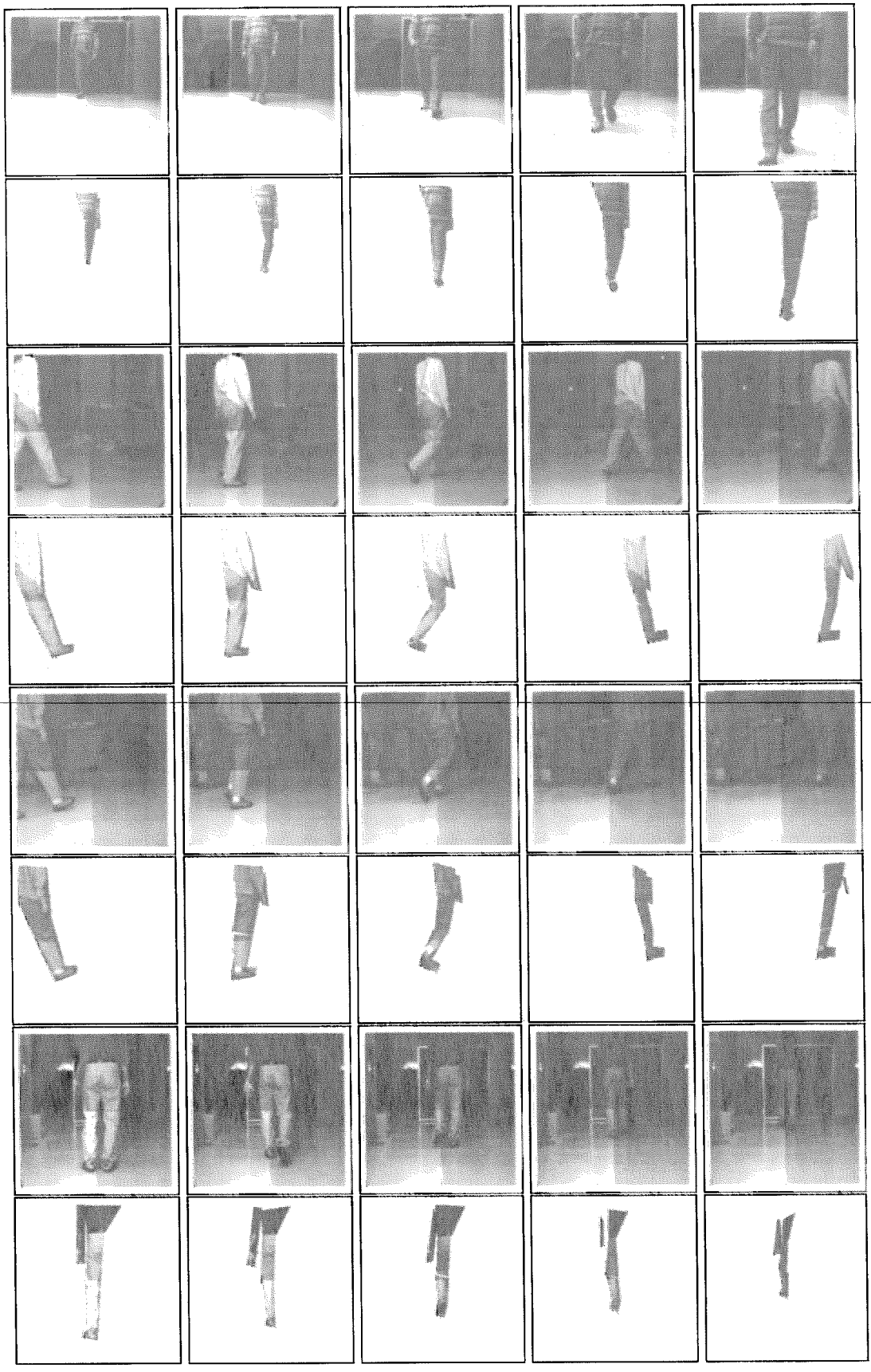

FIG. 8. Tracking examples for four walking directions, forward, diagonal-45 degrees, diagonal-75 degrees, and away (top to bottom, respectively).

TABLE 2
List of Activities and the Number of Occurrence of Each in Training and Recognition

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 3
Confusion Matrix for Recognition Results

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

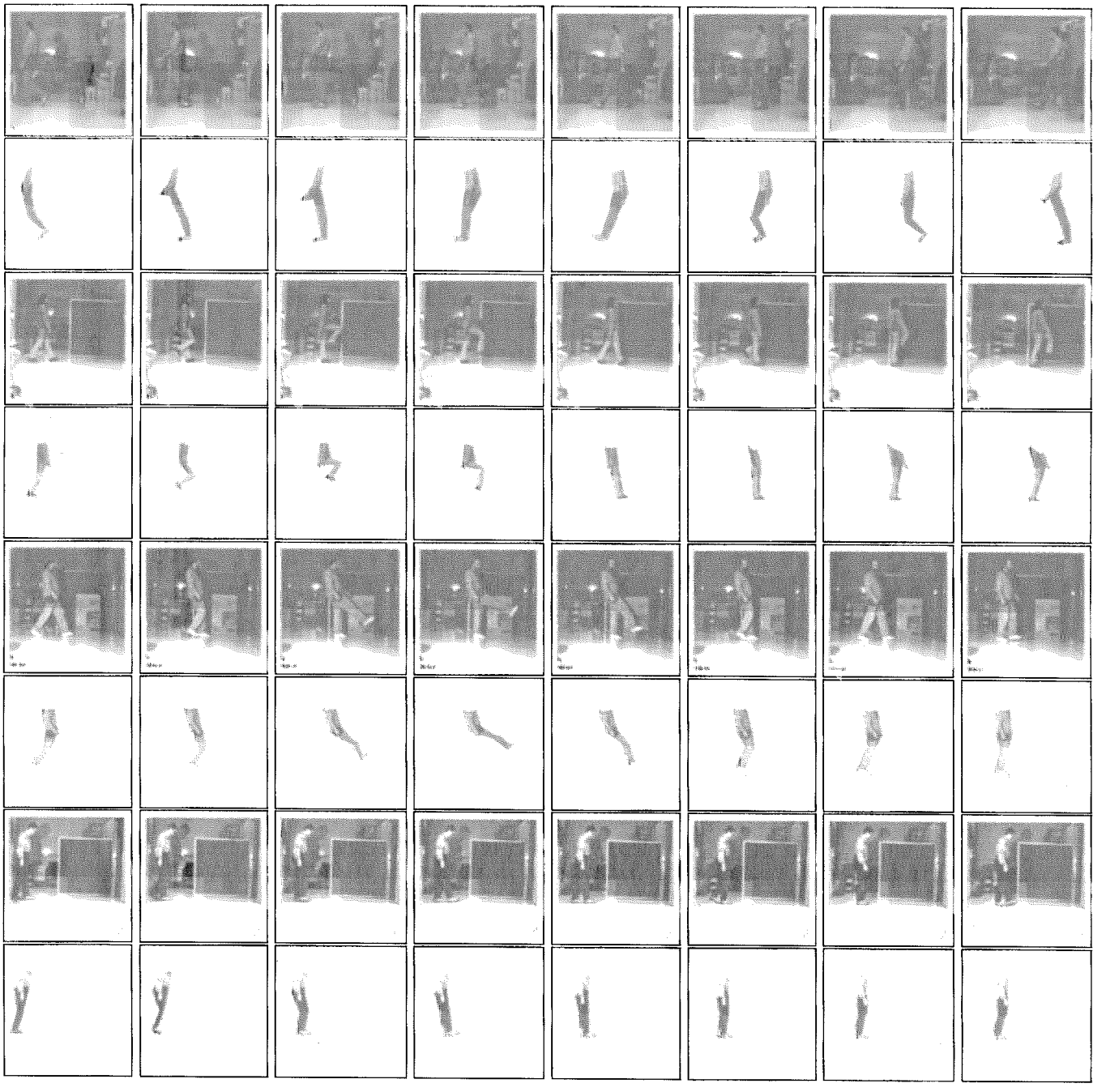

FIG. 9. Frames from image sequences of walking, marching, kicking, and line walking and five part tracking (top to bottom, respectively).

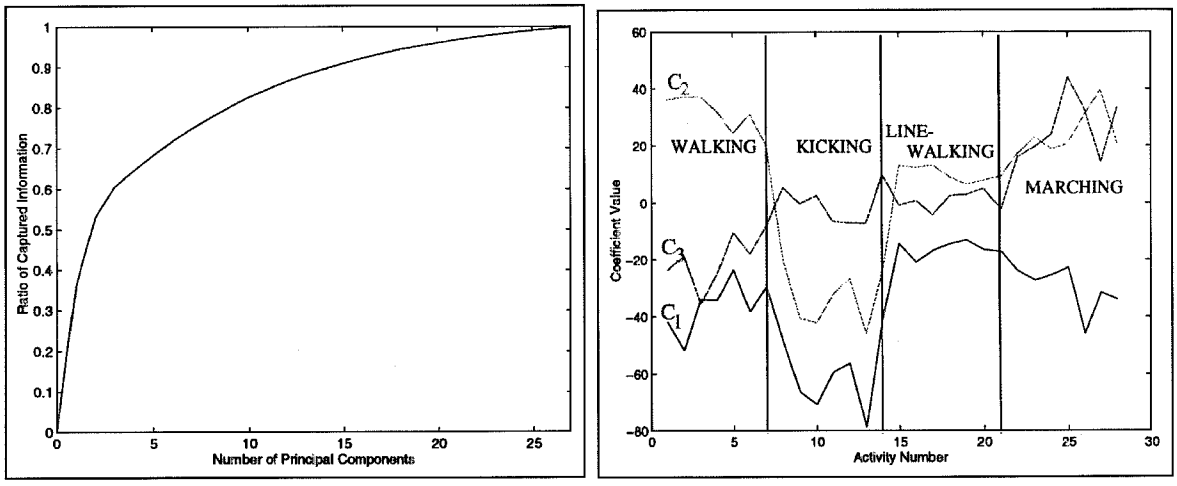

FIG. 10. Cumulative information captured by the 28 basis activities (left) and the expansion coefficients of using the first three activity basis for the 28 activities (right) in which classification among activity is clearly visible (c_1, c_2, and c_3 are shown with the respective delineation of the type of activity in the training set).

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% speedup or slowdown as well as up to 15 frames temporal shift were accounted for by the linear transformation used in the matching.

5.4. *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 (Fig. 11). 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 form the training set of 3120 flow fields and derived a low-dimensional representation using principal component analysis (for a detailed description, see [5] and Appendix).

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 Fig. 12.

Image motion is represented as a linear combination of the basis flow templates: $\sum_{i=1}^{8} m_i M_i(\mathrm{x})$ (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 [5]. 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 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 bases, as shown in Fig. 13. Fifteen activity bases, capture 90% of the temporal variation in these sequences.

Figure 14 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

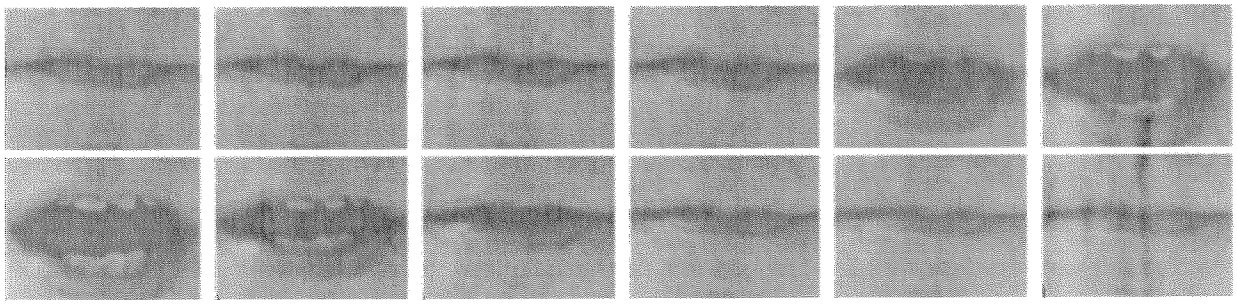

FIG. 11. Example frames for one letter in the training set.

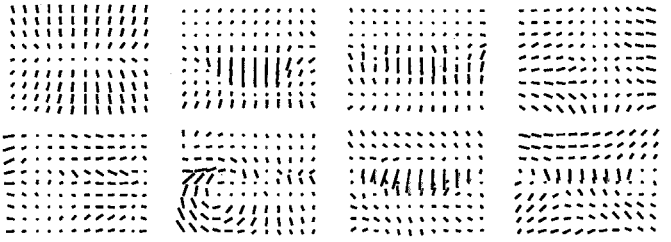

FIG. 12. First eight basis flow fields computed by PCA. They account for 90% of the information in the 793 training flow fields.

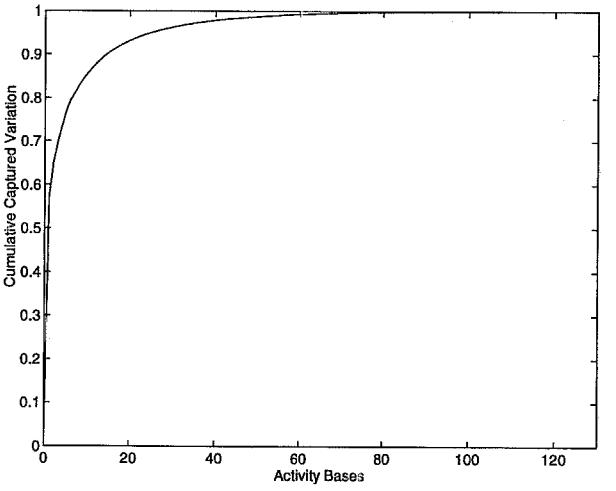

FIG. 13. Cumulative variation captured by 130 basis vectors of the 130 sequences.

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 the intensity variation (see [5]) and use the linear coefficients for 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.

5.5. *Discussion of Experimental Results*

The above experiments have demonstrated the performance of the proposed modeling and recognition approach. The following summarizes our observations on the computational aspects of the algorithm:

- The complexity of the algorithm is a function of the number of free parameters, maximal length of activity bases, and minimization parameters. In the reported experiments the number of estimated parameters (transformation parameters plus the expansion coefficients) has been between 10–20. The length of activity bases has been 160 frames multiplied by the 40 instantaneous parameters for articulated motion. Fifteen iterations of gradient descent were performed. The overall complexity is proportional to $20 * nT * 15$ and is $O(1)$.
- The parameter search may converge to local minima if the initial alignment between the activity models and observed activity is too far to be accounted for by the coarse-to-fine differential formulation of the algorithm. To prevent local minima selection the algorithm is started with several initial alignments and the results are compared so that the global minimum of the

TABLE 4
Confusion Matrix for Recognition of 130 Sequences of 13 Letters

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

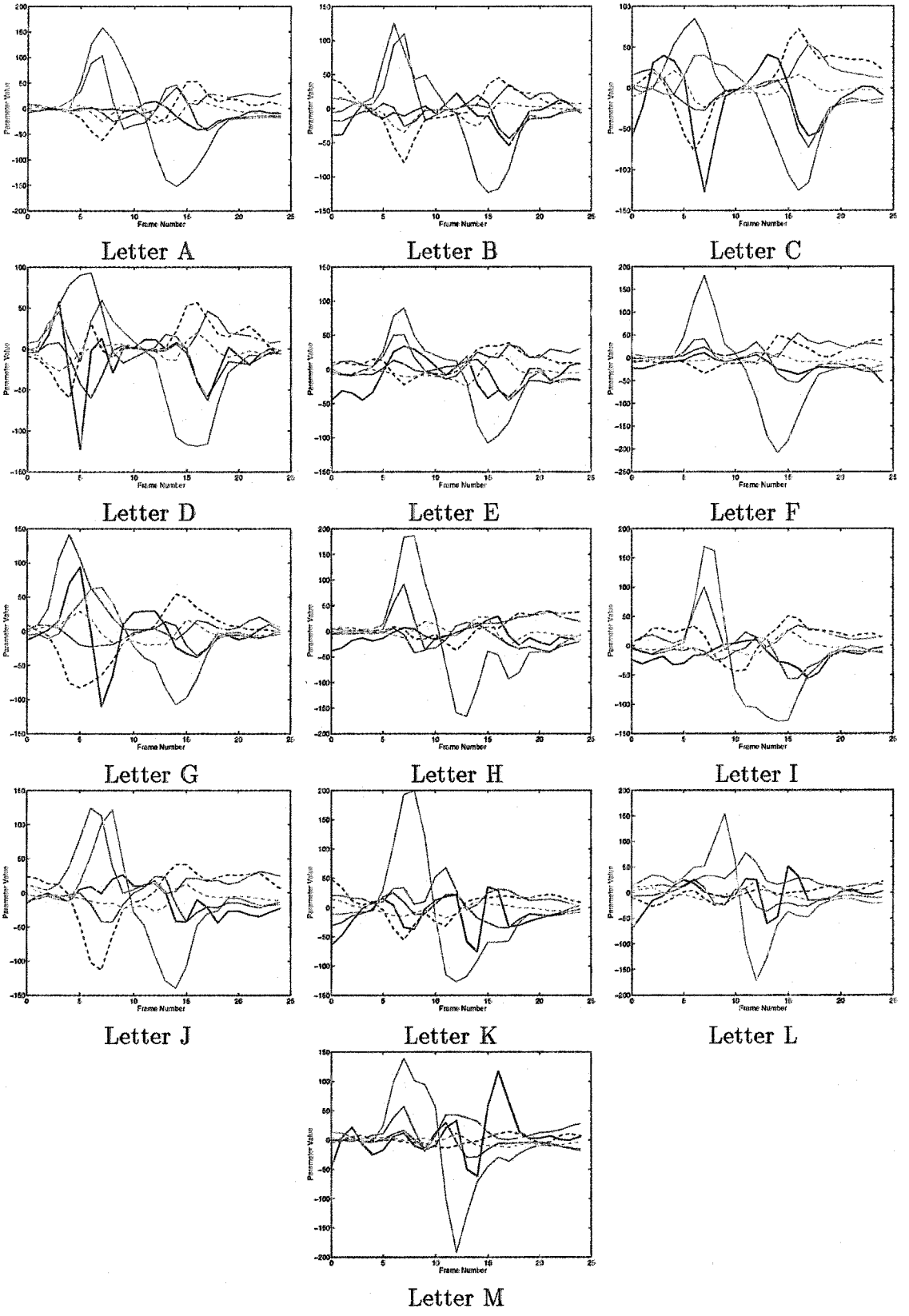

FIG. 14. The eight coefficients of the motion templates computed for each of 13 letters during a complete utterance (using the basis vectors corresponding to the eight largest eigenvalues).

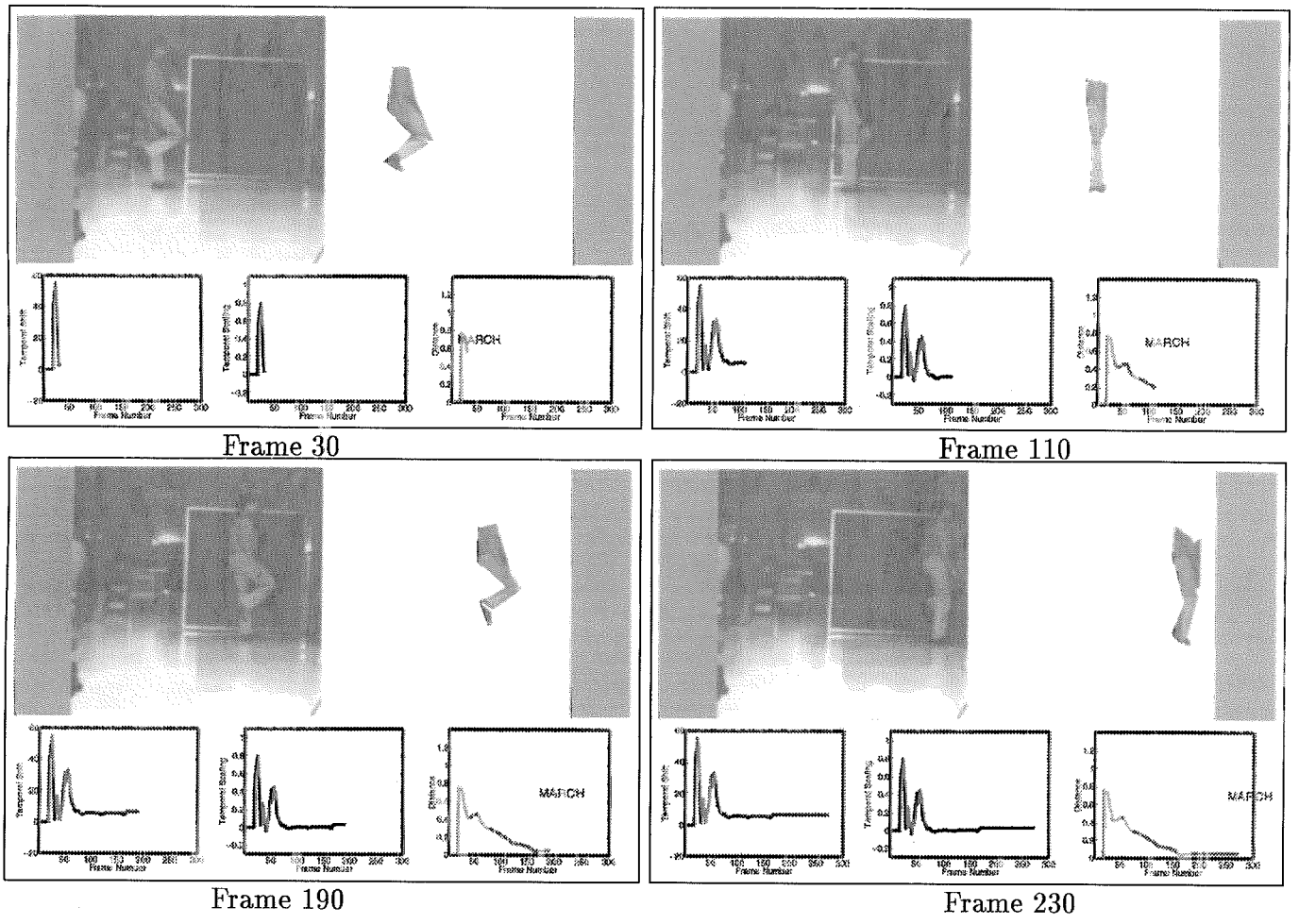

FIG. 15. Tracking and online recognition of a marching activity. Each frame describes the temporal translation, temporal speed, and distance of the observed data with respect to the training examples (left, center, and right, respectively).

error function is chosen. Since the activity model duration is 160 frames for the articulated movement we found that starting with eight points is sufficient. Points are selected at 0, 20, 40, 60, 80, 100, 120, 140 frames relative to the onset of the model.

- The algorithm can be used for online activity recognition. For example, once several tens of frames of an activity have been acquired, recognition can be started and then repeated for each incoming frame. Figure 15 shows a few frames from an online experiment for recognition of marching. The figure shows that a temporal translation of about seven frames and ordinary speed are recovered (bottom left and center graphs, respectively, in each frame). Also it shows for four sample frames the distance of the observed activity from the closest sample in the training set of activities (see the bottom right graph in each frame). This distance is initially large since there are only a few frames of input (see the first 50–70 frames), then it decreases rapidly as more frames are acquired and the approximation of the observed data by the basis activities becomes closer to one of the known activities (an activity from the training set). Eventually this distance goes down to zero if the activity closely resembles a training activity.

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 demonstrated 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 the execution of activities. The formulation we proposed allows future incorporation of nonuniform transformations.

APPENDIX

Articulated Motion Estimation Algorithm

In the following we summarize the articulated motion estimation model proposed in [11]. This model assumes that each body part is a plane moving in a perspective scene. The optical flow of each body patch is given by

$$u(x, y) = a_0 + a_1 x + a_2 y + a_6 x^2 + a_7 xy, \tag{13}$$

$$v(x, y) = a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2, \tag{14}$$

where $\mathbf{a} = [a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7]$ denotes the vector of parameters to be estimated, and $\mathbf{u}(\mathbf{x}, \mathbf{a}) = [u(x, y), v(x, y)]^T$ are the horizontal and vertical components of the flow at image point $\mathbf{x} = (x, y)$. The coordinates (x, y) are defined with respect to the centroid of the whole body region (i.e., a shared center for all parts).

To estimate the motion parameters, $\mathbf{a}_s$, for a given patch, s, we make the assumption that the brightness pattern within the patch remains constant while the patch may deform as specified by the model. This brightness constancy assumption gives rise to the optical flow constraint equation

$$\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_s) + I_t = 0, \quad \forall \mathbf{x} \in \mathcal{R}_s, \tag{15}$$

where $\mathcal{R}_s$ denotes the points in patch s, I is the image brightness function, and t represents time. $\nabla I = [I_x, I_y]$ and I_t are the partial derivatives of image brightness with respect to the spatial dimensions and time at the point $\mathbf{x}$.

For human articulated parts, we assume that each patch is connected to only one preceding patch and one following patch; that is, the patches construct a chain structure. For example, a "thigh" patch may be connected to a preceding "torso" patch and a following "calf" patch. Each patch is represented by its four corners. We simultaneously estimate the motion parameters, $\mathbf{a}_s$, of all the patches. The total error of the motions of the patches (from 0 to n) is

$$E = \sum_{s=0}^{n} E_s = \sum_{s=0}^{n} \sum_{\mathbf{x} \in \mathcal{R}_s} \rho(\nabla I \cdot \mathbf{u}(\mathbf{x}, \mathbf{a}_s) + I_t, \sigma), \tag{16}$$

where ρ is a robust error norm. Since the connected patches motions must agree at the points of attachment, a better constrained equation is given by

$$E = \sum_{s=0}^{n} \left(\frac{1}{|\mathcal{R}_s|} E_s + \lambda \sum_{\mathbf{x} \in \mathcal{A}_s} \|\mathbf{x} + \mathbf{u}(\mathbf{x}, \mathbf{a}_s) - \mathbf{x}' - \mathbf{u}(\mathbf{x}', \mathbf{a}')\|^2 \right), \tag{17}$$

where $|\mathcal{R}_s|$ is the number of pixels in patch s, λ controls relative importance of the two terms, $\mathcal{A}_s$ is the set of articulated points for patch s, $\mathbf{a}'$ is the planar motion of the patch which is connected to patch s at the articulated point $\mathbf{x}$, and $\|\cdot\|$ is the Euclidean norm. The use of a quadratic function for the articulation constraint reflects the assumption that no "outliers" are allowed. The second energy term (the "smoothness" term) in Eq. (17) can also be considered as a spring force energy term between two points. In the examples shown in this paper we track five body parts, thus recovering 40 parameters.

Deformable Motion Estimation Algorithm

In the following we summarize the deformable motion estimation model proposed in [5]. The computation model consists of two components: modeling principal flow templates and estimation of image motion using these templates. Consider the case of mouth motion during speech; we assume that a region of interest, $\mathcal{R}$, has been located and normalized in size to a desired rectangular size (e.g., using the planar face registration in [4]).

The first component consists of two stages. In the first stage the dense optical flow of image sequences with training samples of the mouth motions is computed using [2]. In the second stage a principal component analysis (PCA) of the instantaneous flow fields of the training set of images is computed. The output of this modeling component is a set of q basis flow templates $\vec{m}_i$, $i = 1, \ldots, q$ ($q \ll$ number of input flow fields), each basis vector consists of $2 * n$ elements (n is the number of pixels in the region of interest). The instantaneous flow between any two consecutive images can be well approximated by

$$\vec{f}_k = \sum_{i=1}^{q} c_i \vec{m}_i, \tag{18}$$

where the first n elements represent the horizontal flow and the remaining n elements represent the vertical flow at the n pixels.

The second component formulates an objective function that seeks to best explain brightness movement in a new image sequence using the set of basis flow templates. This is given by

$$E(\vec{c}) = \sum_{\mathcal{R}} \rho\left(\nabla I \cdot \sum_{i=1}^{q} c_i \vec{m}_i \right) + I_t, \sigma \Bigg), \tag{19}$$

where recovery of the c_i coefficients that minimize the error E is performed. Details of the minimization can be found in [5]. The coefficients c_i are the parameters used in the recognition experiments in Section 5.4.

ACKNOWLEDGMENTS

We 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. Yaser Yacoob gratefully acknowledges the support of a DARPA Grant N000149510521.

REFERENCES

1. M. Allmen and C. R. Dyer, Cyclic motion detection using spatiotemporal surfaces and curves, *International Conference on Pattern Recognition* 1990, 365–370.
2. M. Black and P. Anandan, The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Comput. Vision Image Understanding* **63**, 1996, 75–104.
3. M. J. Black and A. Jepson, EigenTracking: Robust matching and tracking of articulated objects using a view-based representation, *IJCV* **26**, 1998, 63–84.
4. M. Black and Y. Yacoob, Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions, *IJCV* **25**, 1997, 23–48.
5. M. J. Black, Y. Yacoob, A. Jepson, and D. Fleet, Learning parameterized models of image motion, *Proc. CVPR, Puerto Rico, June 1997*, pp. 561–567.
6. A. Bobick and A. Wilson, A state-based technique for the summarization and recognition of gesture, *International Conference on Computer Vision 1995*, 382–388.
7. A. Bobick and J. Davis, An appearance-based representation of action, *International Conference on Pattern Recognition* 1996, 307–312.
8. C. Bregler, Learning and recognizing human dynamic in video sequences, *IEEE Conference on Computer Vision and Pattern Recognition* 1997, 568–574.
9. T. Darrell and A. Pentland, Space-time gestures, *Proc. IEEE Conference on Computer Vision and Pattern Recognition 93*, pp. 335–340.
10. 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*, pp. 272–277.
11. 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*, pp. 561–567.
12. N. Li, S. Dettmer, and M. Shah, Visually recognizing speech using eigensequences, in *Motion-Based Recognition* (M. Shah and R. Jain, Eds.), pp. 345–371, Kluwer Academic, Dordrecht, 1997.
13. 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*, pp. 461–465.
14. S. A. Niyogi and E. H. Adelson, Analyzing and recognizing walking figures in XYT, *IEEE Conference on Computer Vision and Pattern Recognition* 1994, 469–474.
15. Polana and R. Nelson, Detecting activities, *IEEE Conference on Computer Vision and Pattern Recognition* 1993, 2–7.
16. K. Rangarajan, W. Allen, and M. Shah, Matching motion trajectories using scale-space, *Pattern Recog.* **26**(4), 1993, 595–610.
17. M. Rosenblum, Y. Yacoob, and L. S. Davis, Human expression recognition from motion using a radial basis function network architecture, *IEEE Trans. Neural Networks* **7**, 1996, 1121–1138.
18. S. M. Seitz and C. R. Dyer, Affine invariant detection of periodic motion. *IEEE Conference on Computer Vision and Pattern Recognition* 1994, 970–975.
19. 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*, pp. 189–194.