SUMMARIZATION AND INDEXING OF HUMAN ACTIVITY SEQUENCES

Bi Song, Namrata Vaswani**, Amit K. Roy-Chowdhury**

*Dept. of EE, University of California, Riverside, CA 92521, {bsong,amitrc}@ee.ucr.edu
**Dept. of ECE, Iowa State University, Ames, IA 50011, namrata@iastate.edu

ABSTRACT

In order to summarize a video consisting of a sequence of different activities, there are three fundamental problems: tracking the objects of interest, detecting the activity change time and recognizing the new activity. This paper presents an algorithm for achieving all these three tasks simultaneously and presents results on how it can used for indexing and summarizing a indexing and summarizing (tracking) a real-life video sequence. Human activities are represented by a model for the dynamics of the shape of the human body contour (shape of k landmarks uniformly chosen on the outer contour). Measures are designed for detecting both gradual transitions and sudden changes between activity models.

1. INTRODUCTION

In order to index and summarize human activity sequences, it is necessary to i) track the activities, ii) detect the change from one activity to the next and ii) recognize the next activity. We develop a novel framework for *persistent and simultaneous* tracking and recognition of human activities consisting of the following steps which take place in a loop: (i) modeling the appearance and motion of single activity sequences and tracking them, (ii) detecting a change from one activity to the next, and (iii) classifying which is the next activity to change to and start tracking it. This paper presents an algorithm for achieving all these three tasks and presents results on how it can used for indexing and summarizing a long sequence consisting of different human activities. Human activities are represented by a model for the dynamics of the shape [1] of the human body contour (shape of k landmarks uniformly chosen on the outer contour). This is motivated by the fact that the shape of the body changes in the course of various activities. Moreover, the shape representation [1] makes the method insensitive to camera zoom (scale changes), translation and in-plane rotation.

Tracking is performed using a particle filter that uses a motion model taken from [2] and a piecewise stationary shape dynamical model [3]. A nonlinear observation equation that relates the predicted landmark configuration with the input image. The piecewise stationary model used here is similar in spirit to switched linear dynamic systems [4], but in our case the state space model is nonlinear. Note that in our framework, the tracked observations are used to recognize an activity, the corresponding dynamical model of which drives the tracking for the next frame. The tracking algorithm is thus similar in spirit to the well-known CONDENSATION algorithm [5], but differs from it in (i) the use of local shape deformation models (as compared to only affine deformation modeling in [5]); (ii) performing simultaneous recognition and tracking using change detection statistics like ELL [6] and tracking error [7]. These measures can handle both slow and sudden changes of activities and serve as a feedback signal, which initiates a search for switching to a new activity model, and the whole process repeats. A diagram explaining our overall approach is shown in Figure 1. We present experimental results on automatically tracking and indexing a real life video sequence of different activities.

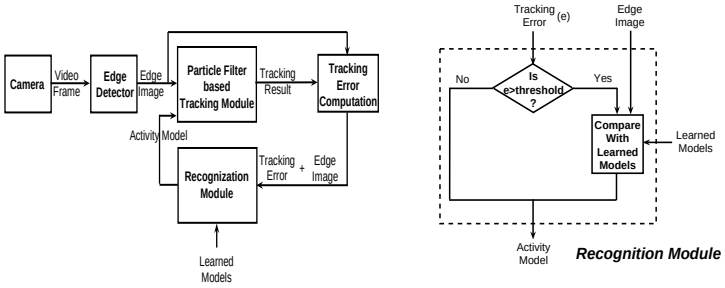

Fig. 1. Overall approach for simultaneous tracking and recognition. The recognition module for using Tracking Error is shown on the right. An analogous module operates for using ELL as well.

1.1. Relation to Previous Work

There has been much recent work on human activity recognition [8, 9, 10, 11]. Key-frame segmentation methods [12] only detect the switching instances and often require the entire video to be available a-priori. Video surveillance methods also address the problems of tracking and recognition, but usually the tracks are obtained first, followed by recognition [13, 2].

Simultaneous tracking of the moving persons and recognition of activities has been performed in many applications using a Dynamic Bayesian Network (DBN) model tracked by a Rao-Blackwellized particle filter [14, 15, 4, 16]. The events are recognized by tracking discrete state space variables (whose dynamics is defined by the DBN) using a particle filter (PF) and moving object motion is modeled by a linear dynamical system tracked by a Kalman filter inside the PF. [4] performs figure tracking by defining a DBN to switch between various linear dynamical systems (also called Switched Linear Dynamical System or SLDS). Discrete state space variables have also been used in many other joint recognition and tracking contexts in video analysis, e.g. Chapter 16 of [15] (Condensation for gesture tracking and recognition), [17, 18].

Using a discrete mode as a state variable requires knowledge of its dynamics. In cases when this is not known, one can choose to detect a change by using tracking error (TE) [7] or the recently proposed ELL statistic [6] and then recognize the new activity by matching it with the activities in the database. TE based model switching and re-initialization [7] is a common technique in systems and control literature. But in computer vision it has been used with only limited success because of the difficulty of re-initialization. In the current application, we are able to do this successfully. Also, for

gradual changes, we are able to prevent large loss of track from occurring by using ELL which is able to detect a change before complete loss of track.

We start by describing the state space model (Section 2), followed by tracking using filters, change detection and recognition strategy (Section 3). We then show detailed experiments and analyze the results (Section 4).

2. STATE SPACE MODEL FOR ACTIVITY SEQUENCES

Our state space is the shape and global motion (scale, rotation, translation) of k landmark points used to represent the outer contour of the object of interest. In past work [2, 3], we have extended the statistical theory for landmark shapes [1] to define stochastic dynamic models for shape deformation. We model the motion/deformation of a changing configuration of landmark points as scaled Euclidean motion (translation, rotation, isotropic scaling) of a "mean shape" plus its non-rigid deformation. The term "shape activity" is used to denote a particular stochastic model for shape deformation. We define a "stationary shape activity" (SSA) as one for which the mean shape remains constant with time and the deformation model is stationary [2]. A piecewise stationary shape activity (PSSA) model [3] models a shape activity with slowly varying "mean shape" (approximated as piecewise constant). It is represented by a sequence of SSAs with nonstationary transitions which we detect using ELL [6] or tracking error [7].

The state vector $X_t = [v_t, s_t, \theta_t, a_t, b_t]$ where $v_t = v(z_t, \mu)$ denotes the tangent coordinates [1] of the shape, z_t, computed in the tangent space of μ and s_t, θ_t, a_t, b_t denote the isotropic scale, rotation, x and y translation. Complex notation taken from [1] is used simplify writing of equations. * denotes conjugate transpose of a complex vector and $j = \sqrt{-1}$. The predicted configuration of landmark points at time t is $h(X_t) = z_t s_t e^{j\theta_t} + a_t + jb_t$ where the shape, $z_t = (1 - v_t^* v_t)^{1/2}\mu + v_t$.

We use the same global motion model as in [2]. We describe the shape dynamical model below.

2.1. Piecewise Stationary Shape Activity (PSSA) Model

We refer the reader to [3] for more details about the PSSA model. Let the "mean shape" change times be $t_{\mu_1}, t_{\mu_2}, t_{\mu_3}, ...$ and the corresponding means be $\mu_1, \mu_2, \mu_3, ...$ Between $t_{\mu_{j-1}} \leq t < t_{\mu_j}$, $\mu_t = \mu_{j-1}$ and thus $v_t = v_t(z_t, \mu_{j-1})$. During this interval, the dynamics is similar to that for an SSA, i.e.,

$$\begin{aligned} v_t(z_t, \mu_{j-1}) &= A_v v_{t-1}(z_{t-1}, \mu_{j-1}) + n_t,\ n_t \sim \mathcal{N}(0, \Sigma_{v,t}) \\ z_t &= (1 - v_t^* v_t)^{1/2}\mu_{j-1} + v_t. \end{aligned} \quad (1)$$

At the change time instant, $t = t_{\mu_j}$, $\mu_t = \mu_j$ and so the tangent coordinate v_{t-1} needs to be recalculated in the new tangent space with respect to $\mu_t = \mu_j$. This is achieved as follows [3]:

$$v_{t-1}(z_{t-1}, \mu_j) = [I - \mu_j \mu_j^*] z_{t-1} e^{j\ \angle z_{t-1}^* \mu_j)} \quad (2)$$

Once this is done, the equations of (1) apply with mean shape μ_j.

2.2. Observation Model

We perform edge detection on the image I_t and use the edge map, $G_t = \Upsilon(I_t)$, to obtain the observed landmarks, $\Gamma_t \subset G_t$. Our method is inspired by [5]. Given the predicted location of landmarks, $\hat{Y}_t = h(X_t) = z_t s_t e^{j\theta_t} + a_t + jb_t$, we search along the normal direction to each predicted landmark until we find an edge point and we treat this as the observed landmark location. Thus the observation likelihood is

$$p(\Gamma_t | X_t) \propto exp\{-\sum_{k=1}^{K} \frac{1}{2r_k K} ||q_k - f(q_k, G_t)||^2\}, \quad (3)$$

where K is the shape vector dimension, r_k is the variance of the k^{th} landmark, q_k is the k^{th} predicted landmark, i.e., $q_k = \hat{Y}_{t,k}$ and $f(q_k, G_t) = \Gamma_t$ is the nearest edge point of q_k along its normal direction.

3. TRACKING, CHANGE DETECTION, RECOGNITION

3.1. Tracking using Particle Filters

In this paper, we use a particle filter for "tracking", i.e., for obtaining observations on the fly by tracing along the normals of the predicted configuration, $\hat{Y}_t$, to search for the closest edge (as described in Section 2.2). The particle filter (PF) is a sequential Monte Carlo method (sequential importance sampling plus resampling) which provides at each t, an N sample Monte Carlo approximation to the prediction distribution, $\pi_{t|t-1}(dx) = Pr(X_t \in dx | Y_{1:t-1})$, which is used to search for new observed landmarks. These are then used to update $\pi_{t|t-1}$ to get the filtering (posterior) distribution, $\pi_{t|t}(dx) = Pr(X_t \in dx | Y_{1:t})$. We use a particle filter because the observation model is nonlinear and the posterior can temporarily become multimodel when there are false edges due to background clutter.

3.2. Change Point Detection

As explained earlier, each activity is represented by an SSA or a PSSA (sequence of SSAs) model, for example the bending across activity shown in Figure 2 is composed of 3 SSA pieces. The sequence of activities forms a long PSSA. We use ELL [6] and tracking error [7] described below to detect the change time from one SSA to the next. If the change is gradual as, for example within an activity (e.g. see bending across activity and the ELL plot in Figure 2), the loss of track is small and slow. For such examples, ELL detects the change faster than tracking error. For our state space model, ELL is computed as

$$ELL_t^N = \frac{1}{N} \sum_{i=1}^{N} {v_t^{(i)}}^T \Sigma_v^{-1} v_t^{(i)} + constant \quad (4)$$

where N is the number of particles and Σ_v is the covariance matrix of the tangent coordinates for the current stationary piece.

If the activity change is sudden, it will cause the PF, and tuned to the dynamical model of a particular activity, to lose track when the activity changes. This is because under the existing activity model with which the PF operates, the new observations would appear to have very large observation noise. Thus the tracking error (TE) [7] (Euclidian norm of the error between the mean predicted landmark configuration and the observed one) will increase when the activity changes and this can be used to detect the change times.

3.3. Model Switching to a New Activity

Once the change has been detected, the next problem is to determine the correct activity from the class of previously learned activity models. This is known as the problem of *model switching*. This is done by projecting the observed landmark configuration, Γ_t, onto

the mean shape for each of the learned activities and choosing the one with the largest projection or the smallest projection error (measured using Procrustes distance [1]). In practice, this is done for a few frames before a final decision is made, since individual frames of different activities may be similar. If the distance is above a certain threshold for all activities in the database, we decide that the current activity is not within the learned database and this is indicated.

3.4. Simultaneous Tracking, Change Detection and Recognition (Simul-TraCR) Algorithm

We now outline the main steps of the simultaneous tracking and recognition algorithm, incorporating change detection and model switching. For simplicity, let us assume that there are two activities in the sequence, A_1 and A_2. For the first frame in A_1, the region of interest (a person or a group of people) is detected based on the application requirements (not part of this paper) and the corresponding model for the activity is determined as in Section 3.3. After this initialization, the algorithm now proceeds as follows.

Track Based on the detected region and the chosen dynamical model, the particle filter is used to track the activity. Measures for determining the accuracy of the tracking algorithm (TE and ELL) are computed for each frame.

Change Detection When the fidelity measures exceed a certain threshold (details in Section 4) for a few consecutive frames, a change is detected.

Model Switching Once the change is detected, the new shape vector is obtained from the edge map of image frame and a search is initiated for the correct activity model. Once the correct activity model is identified, we use this and go back to Track.

Note that change detection and switching may be between different portions of the same activity, specifically, for those activities in which a non-stationary dynamical model is needed.

4. EXPERIMENTAL RESULTS

We now show examples of our Simul-TraCR algorithm for indexing and summarizing (tracking) a sequence consisting of 10 different activities captured in video. The training and testing sequences were captured separately on different days. The binarized silhouette denoting the contour of the person in every frame of the *training* sequence is obtained using background subtraction. The landmarks were obtained by uniformly sampling the silhouette contour. The global motion and shape is computed for the landmark configuration at each frame [1]. This is used to learn the parameters of the dynamical model for each SSA activity as discussed in [2]. In the *testing* sequence, the silhouette is pre-computed only in the first frame if the background information is available; otherwise we use motion segmentation over a few initial frames to obtain the silhouette. Thereafter it is obtained as the output of the tracking algorithm, as explained above. The database we collected consists of 10 activities (whose composition make up a number of normal everyday activities), bending across, walking towards camera and bending down, leaning forward and backward, leaning sideward, looking around, turning head, turning upper body, squatting, bending with hands outstretched, and walking. We will refer to the n^{th} activity as Act$n$.

In Figure 2, we show four frames from the 3 stationary pieces (SSAs) that constitute Act1 (Bending Across) and the ELL and tracking error plots. The first row shows one frame from each piece - Standing Straight (SS), Half Bent (HB) and Fully Bent (FB). The transitions SS-HB and HB-FB were gradual and hence are detected by ELL faster than by Tracking Error. The pink horizontal lines in the ELL plot are the average value of ELL for that activity piece (equal to the effective rank of Σ_v) and change is declared when ELL significantly exceeds this value. ELL is always computed w.r.t. the SSA that is being used to track the current frame. ELL detects change before significant loss of track and hence we switch to the next SSA piece without the tracking error ever increases appreciably (varies about an average value of about 60). The first image in the bottom row is frame 54 .As can be seen the landmarks on the left arm are very close to each other and overlapping (change in topology of underlying continuous contour). Thus there is a change in their order. Landmark shape is sensitive to the ordering of landmarks and this is detected as a sudden increase in ELL.

In Figure 3, we show some frames from a set of individual activities stitched together and the Tracking Error plot. Since the activities were stitched together by us, the transitions from one to the next are sudden. This models a situation where disparate activity videos (i.e., not a continuous sequence) are stitched together, like in a digital library. These are detected easily using the increase in Tracking Error. The plot is for the following sequence: Act3, Act4, Act8, Act9 and Act7. One frame for each activity along with the tracking error is also shown. The number of frames that are used to recognize an activity is called the "delay" due to model switching. The following observations were made in the experimentation process. For Act7, Act8 and Act9, the delay needed to get correct recognition will be very small, while Act3 and Act4 need longer delays to find the correct model to switch to. This is because initial poses of the body in Act3 and Act4 is very similar to other activities.

5. CONCLUSIONS

In this paper, we proposed a novel system for indexing and summarizing (tracking object of interest) a video consisting of a sequence of human activities. This is achieved through an algorithm for simultaneous and persistent tracking and recognition. We use a non-linear, piecewise stationary model defined on the shape of human body contour to represent activities. The activity change times are detected using ELL and Tracking Error statistics. The activities are recognized by comparing the tracked observations against a prior database. We demonstrate the effectiveness of our system by showing experimental results on real life video of different activities.

6. REFERENCES

[1] I. Dryden and K. Mardia, *Statistical Shape Analysis*, John Wiley and Sons, 1998.

[2] N. Vaswani, A. Roy-Chowdhury, and R. Chellappa, "Shape Activities: A Continuous State HMM for Moving/Deforming Shapes with Application to Abnormal Activity Detection," *IEEE Trans. on Image Processing*, October 2005.

[3] N. Vaswani and R. Chellappa, "NonStationary Shape Activities," in *Proc. of IEEE Conf. on Decision and Control*, 2005.

[4] T.-J. Cham and J. M. Rehg, "A multiple hypothesis approach to figure tracking," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*.

[5] M. Isard and A. Blake, "Condensation: Conditional Density Propagation for Visual Tracking," *International Journal of Computer Vision*, pp. 5–28, 1998.

[6] N. Vaswani, "Change Detection in Partially Observed Nonlinear Dynamic Systems with Unknown Change Parameters," in *American Control Conference*, 2004.

[7] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*, Academic Press, 1988.

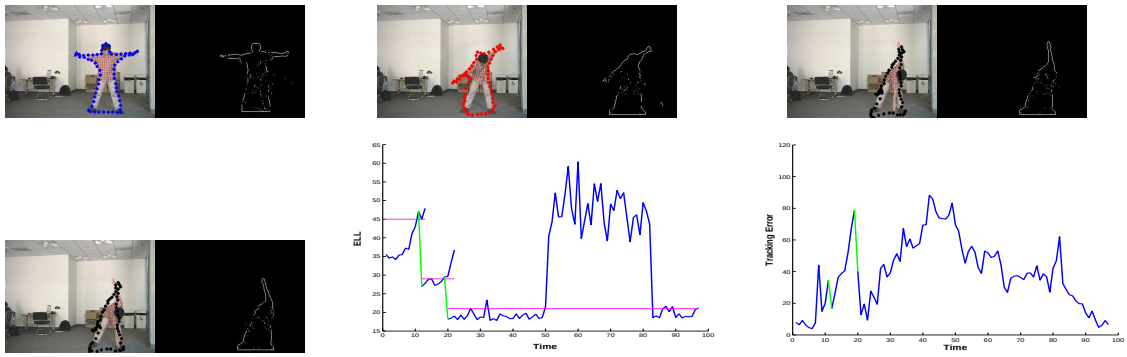

Fig. 2. Tracking Act1 (Bending Across). The top row has one frame from each SSA piece of this activity - Standing Straight (SS), Half Bent (HB) and Fully Bent (FB). Bottom row, first image is another FB frame where there is change in the ordering of landmarks near the arm (detected by the increase in ELL after frame 50). Second image is the ELL plot. ELL detects the gradual transitions SS-HB and HB-FB. The last image is the Tracking Error plot. ELL detects the change before large loss of track and we switch to the next model. Hence Tracking Error never increases too much.

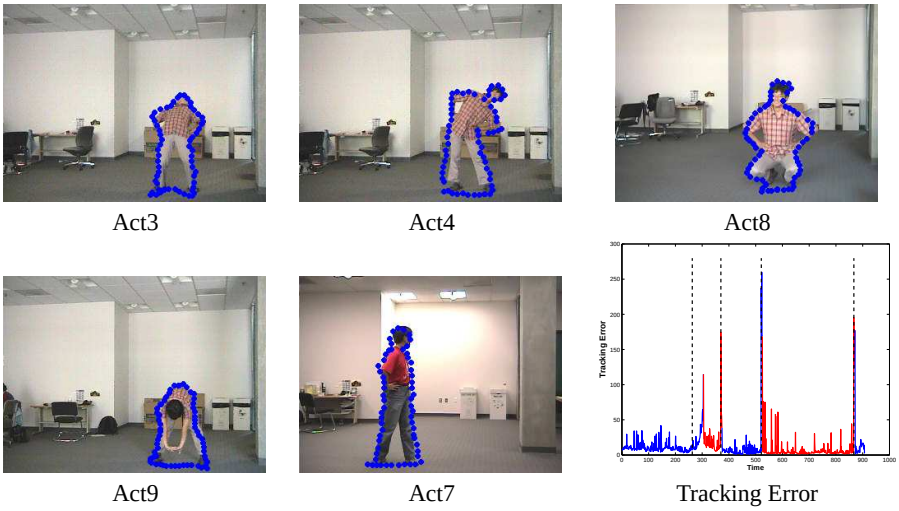

Fig. 3. One frame for each activity along with superimposed tracking error is shown. Tracking error to detect switches in a multi-activity sequence. The switches between activities are sudden. This models a situation where disparate activity videos (i.e., not a continuous sequence) are stitched together, like in a digital library. The tracking error increases when an activity switch happens. Once the model switch occurs and the new model is able to track properly, the tracking error goes down

[8] N. Lobo P. Smith, M. Shah, "Integrating and employing multiple levels of zoom for activity recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[9] L. Zelnik-Manor and M. Irani, "Temporal factorization vs. spatial factorization," in *Proc. of European Conference on Computer Vision*, 2004.

[10] S. M. Khan and M. Shah, "Detecting group activities using rigidity of formation," in *ACM Multimedia*, 2005.

[11] D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, January 1999.

[12] Y. Zhai and M. Shah, "A general framework for temporal video scene segmentation," in *Proc. of International Conf. on Computer Vision*, 2005.

[13] W.E.L. Grimson, L. Lee, R. Romano, and C. Stauffer, "Using Adaptive Tracking to Classify and Monitor Activities in a Site," in *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, 1998, pp. 22–31.

[14] L. Liao, D. Fox, and H. Kautz, "Location-based activity recognition using relational markov networks," in *Proc. of the International Joint Conference on Artificial Intelligence*, 2005.

[15] A. Doucet, N.de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

[16] D. Wilson and C. Atkeson, "Simultaneous Tracking and Activity Recognition (STAR) Using Many Anonymous, Binary Sensors," in *Proceedings of PERVASIVE*, 2005.

[17] S.K. Zhou, R. Chellappa, and B. Moghaddam, "Visual Tracking and Recognition Using Appearance-Adaptive Models in Particle Filters," *IEEE Trans. on Image Processing*, vol. 13, no. 11, pp. 1491–1506, November 2004.

[18] M. Harville and D. Li, "Fast, integrated person tracking and activity recognition with plan-view templates from a single stereo camera," in *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, 2004, pp. II: 398–405.