Closed-Loop Tracking and Change Detection in Multi-Activity Sequences *

Bi Song [1], Namrata Vaswani [2], Amit K. Roy-Chowdhury [1]
[1]Dept. of Electrical Engineering, University of California, Riverside, CA 92521
{bsong,amitrc}@ee.ucr.edu
[2]Dept. of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011
namrata@iastate.edu

Abstract

We present a novel framework for tracking of a long sequence of human activities, including the time instances of change from one activity to the next, using a closed-loop, non-linear dynamical feedback system. A composite feature vector describing the shape, color and motion of the objects, and a non-linear, piecewise stationary, stochastic dynamical model describing its spatio-temporal evolution, are used for tracking. The tracking error or expected log likelihood, which serves as a feedback signal, is used to automatically detect changes and switch between activities happening one after another in a long video sequence. Whenever a change is detected, the tracker is reinitialized automatically by comparing the input image with learned models of the activities. Unlike some other approaches that can track a sequence of activities, we do not need to know the transition probabilities between the activities, which can be difficult to estimate in many application scenarios. We demonstrate the effectiveness of the method on multiple indoor and outdoor real-life videos and analyze its performance.

1. Introduction

Video sequences often have people performing multiple activities one after another. For example, Fig. 1 shows some frames from an educational Yoga video where the instructor transits continuously between different postures. In multi-camera environments spanning a large geographical area, it is common for people to do different activities which we need to continuously track. A recent review [8] suggested that tracking long sequences of activities is still a challenging problem. This motivates us to design methods that can track a long multi-activity sequence including the time instances of change from one activity to the next, using a closed-loop, non-linear dynamical feedback system.

*The authors at UCR were partially supported by NSF Grants 0551741 and 0622176.

1.1. Overview of Proposed Method

Our framework for tracking of human activities consists 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 sequence to the next, and (iii) automatically reinitializing after the change and starting tracking. A diagram explaining our overall approach is shown in Fig. 1.

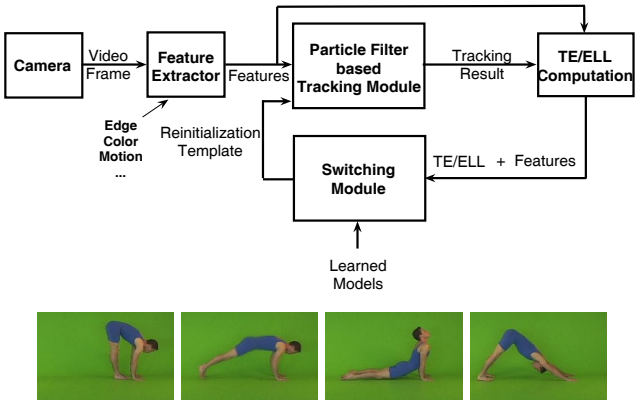


Figure 1. Framework for closed-loop tracking and an application scenario.

Activities are represented mathematically using dynamical models defined on the shape of the contour of the human body, motivated by the fact that the shape of the body changes in the course of various activities. The shape representation makes the method insensitive to camera scale changes and translation. Tracking is performed using an auxiliary particle filter [16], incorporating piecewise stationary dynamical models and a nonlinear observation equation that relates the shape, color and motion with the input image. The reason for choosing a particle filter over an extended Kalman filter (EKF) is multimodal observation likelihood due to either background clutter or partial occlusions. The estimated tracking error (TE) [1] or expected log likelihood (ELL) [19]) are used to detect when the activity deviates from the current model.

1-4244-1180-7/07/$25.00 ©2007 IEEE

We start by placing our work in the context of existing literature. Then we describe the theoretical background of our framework, including state space modeling of human activities (Section 2) and our observation model (Section 3), followed by the tracking and change detection algorithm (Section 4). We then show detailed experiments and analyze the results (Section 5).

1.2. Relation to Previous Work and Contributions

A review of recent work in human activity analysis can be found in [8]. Based on the conclusions of [8] and a survey of some of recent methods on event analysis (see major journals and conference proceedings), we find that most existing methods handle events of short duration with moderate changes in illumination, scene clutter and occlusion. The main *contributions* of this work are :
(i) tracking in a *long* video sequence consisting of multiple activities happening sequentially in changing background, occlusion and clutter,
(ii) achieving the above without prior knowledge of the transition probabilities between different activities.

A few techniques have studied the problem of tracking multi-activity sequences using the idea of simultaneous tracking and recognition. In [13], the idea of integrated tracking and recognition of human activities using particle filters was proposed. Simultaneous tracking of moving persons and recognition of their activities has also been performed in many applications using a Dynamic Bayesian Network (DBN) model [2, 4, 12, 23]. However, all these methods require a-priori knowledge of the transition probability matrix for switching between different activity models. Designing such a transition matrix for uncontrolled environments like video surveillance may be difficult (since the class of possible activities is large and can change arbitrarily from one to another) and the process has to repeat for each application scenario.

In contrast to this open-loop approach, we propose to use change detection measures (e.g., tracking error/ELL) to detect transitions between activities, and use these as a *feedback signal* in a closed-loop system. The main advantage of our approach is significant computational saving in tracking changes of body posture using the particle filtering framework. Also we do not need to learn the transition probabilities between the various activities.

For modeling activities, we use a discrete shape representation which is different from level set representations of shapes such as those described in [10, 17]. The level set approach is theoretically infinite (and in practice large time varying finite) dimensional, and hence defining dynamics on and sampling from such a large dimensional space is computationally expensive. This is overcome through the use of various approximate tracking solutions, as in [10, 17]. Level sets, however, have the advantage that they can adjust to large changes in the shape and topology, which is usually problematic for discrete representations. For large changes in shape, we show that it is possible to overcome this problem for many activity recognition applications by using piecewise stationary dynamical models. Moreover, a discrete representation allows adoption of the overall framework to different descriptions of human body structure (e.g., sampled contours, joint angles, stick figure models, etc.). We describe our approach on the sampled contour (similar to [21]), which can be easily replaced by other features on the image.

2. Modeling of Human Activities

As mentioned earlier, we use a composite feature vector comprised of shape, motion and color for tracking. We now describe the shape feature and associated dynamical models. The integration of color is described in the next section.

We model human activities by the shape of the contour of the body and its motion (rigid and nonrigid) over time. We model the motion/deformation of a deforming shape as scaled Euclidean motion (i.e., translation, rotation, isotropic scaling) of a "mean shape" plus its non-rigid deformation [24]. The term "shape activity" [21] 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. A piecewise stationary shape activity (PSSA) model [20] 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 the TE or ELL. Fig. 2 shows an example.

2.1. Discrete Shape-Based Representation

We use a discrete representation of the shape, z_t^k, for a group of k landmarks on the body contour at time t. Since the shape z_t^k lies on the complex spherical manifold $\mathcal{C}^{k-1}$ [5], we denote its tangent coordinates at a mean shape μ as $v_t(z_t, \mu) = [I_k - \mu\mu^T]z_t$ (we drop k for notational simplicity) [5]. Let s_t, θ_t, a_t, b_t denote the isotropic scale, rotation, and x and y translation, respectively, in normalizing from the collection of landmarks to the shape space. Let us now define a state vector $X_t = [v_t, s_t, \theta_t, a_t, b_t]$. The configuration of landmark points is related to the shape as $h(X_t) = z_t s_t e^{j\theta_t} + a_t + jb_t$.

2.2. Dynamical Model of Human Activities

The shape dynamics consist of the deformation and global motion of the body contour. When the shape is not stationary but is slowly varying, one could model the mean shape as being piecewise constant (see Fig. 2). These mean shapes are similar to templates or exemplars used in activity

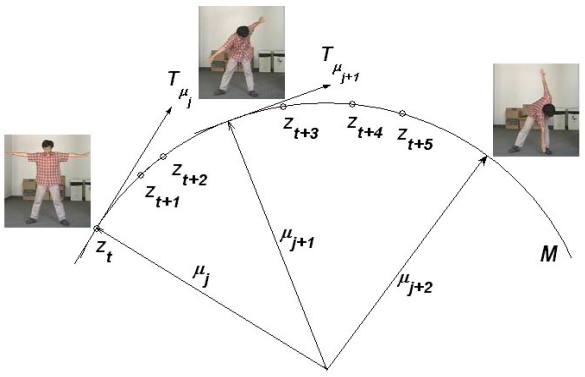


Figure 2. Piecewise-Stationary Shape Sequences on the shape manifold which is depicted using a circle ($\mathcal{M}$), instead of a complex $\mathcal{C}^{k-1}$ sphere.

recognition algorithms, e.g., the exemplars in HMM based approaches [11, 22], with the difference that we work in shape space.

Let the mean shape change times be $t_{\mu_1}, t_{\mu_2}, t_{\mu_3}, ...$ (see Fig. 2) and the corresponding means be $\mu_1, \mu_2, \mu_3, ...$ Then we have the following dynamics for the *deformation*: between $t_{\mu_{j-1}} \leq t < t_{\mu_j}$, $\mu_t = \mu_{j-1}$, the dynamics is modeled by a stationary Gauss-Markov process, i.e.,

$$v_t(z_t, \mu_{j-1}) = v_{t-1}(z_{t-1}, \mu_{j-1}) + n_t, n_t \sim \mathcal{N}(0, \Sigma_n). \quad (1)$$

At the change time instant, $t = t_{\mu_j}$, $\mu_t = \mu_j$ and so the tangent coordinate v_t needs to be recalculated in the new tangent space with respect to $\mu_t = \mu_j$, as [20]

$$v_t(z_t, \mu_j) = [I_k - \mu_j {\mu_j}^T] z_t e^{j\theta(z_t, \mu_j)}. \quad (2)$$

To represent the Euclidean motion of the mean shape, we use a Gauss-Markov model for log-scale, $\log s_t$, and a Markov uniform model for θ_t [21]. More details on the PSSA model are available in [20].

Training: Given a training sequence of centered (translation normalized) configurations, $\{Y_t\}_{t=1}^T$, for a particular stationary activity sub-model , the mean shape μ_{j-1} is estimated by Procrustes mean [5] of $Y_t, t_{\mu_{j-1}} \leq t < t_{\mu_j}$. The scale s_t is the Euclidean norm of Y_t. The rotation angle is given by $\theta_t(Y_t, \mu_{j-1}) = -angle((Y_t/s_t)^T \mu_{j-1})$. The shape, z_t, which is defined on configuration, Y_t, and mean shape, μ_{j-1}, is $z_t(Y_t, \mu_{j-1}) = (Y_t/s_t)e^{j\theta_t}$. Then the tangent coordinates of the shape is [5]

$$v_t(z_t, \mu_{j-1}) = [I_k - \mu_{j-1}\mu_{j-1}^T] z_t. \quad (3)$$

Assuming a time invariant Markov model on v_t, we can use $\{v_t\}_{t=1}^T$ to learn its parameters as [15]

$$\Sigma_n = \frac{1}{T} \sum_{t=1}^{T} (v_t - v_{t-1})(v_t - v_{t-1})^T. \quad (4)$$

This is similar to the Active Shape Models (ASM) approach [3], except that activities are represented in shape space.

3. Observation Model for Tracking

The observation process is defined by the likelihood distribution, $p(I_t|X_t)$, where X_t is the state vector and I_t is the image observation at t. Our observation models were generated by combining a model of shape with a model of the appearance.

The shape representation requires the availability of the landmark points. We will use the edge image using a Canny detector to obtain the landmarks. Our shape observation is the edge image, $G_t = \Upsilon(I_t)$, (where Υ denotes the edge extraction operator) . The observation likelihood describes the probability of a set of landmark points, $\Gamma_t \subset G_t$, given the predicted state vector, $X_t = [v_t, s_t, \theta_t, a_t, b_t]$. Let $\hat{Y}_t = h(X_t) = s_t z_t e^{-j\theta_t}$ be the predicted configuration of landmarks based on the dynamical model. It is assumed that a mapping, f, is known that associates each predicted landmark of $\hat{Y}_t$ with a point on the edges. In practice this mapping is set up by searching for the closest edge along the normal of the predicted configuration and this is treated as the observed landmark, Γ_t. Thus the shape observation likelihood is [9]

$$p_s(\Gamma_t|X_t) = \gamma_s \exp\{-\sum_{l=1}^{k} ||q_l - f(q_l, G_t)||^2\}, \quad (5)$$

where k is the shape vector dimension, q_l is the l^{th} predicted landmark, i.e., $q_l = \hat{Y}_{t,l}$, $f(q_l, G_t) = \Gamma_t$ is the nearest edge point of q_l along its norm direction and γ_s is normalizing factor.

To build a color appearance model we sample the color information from the input image I_t over the region covered by the predicted configuration $\hat{Y}_t$. In order to get rid of the effect of illumination, we use the comprehensive color normalization (CCN) algorithm [6] to get the normalized color image first. For each predicted landmark, l, of $\hat{Y}_t$, we sample a set of a certain number of pixels, $\{o_n, n = 1...N\}_l$, in an elliptical region whose major axis is aligned with the local normal at l. The color histogram associated with each landmark of $\hat{Y}_t$, ch_l, is calculated over $\{o_n, n = 1...N\}_l$.The appearance observation likelihood is defined as [14]

$$p_a(I_t|X_t) = \gamma_a \exp\{-\sum_{l=1}^{k} ||B(ch_l, CH_l)||^2\}, \quad (6)$$

where k is the shape vector dimension, $B(.)$ is the Bhattachayya distance between two color histograms, CH_l is the color histogram associated with the l^{th} landmark of Y_0 (the initialized configuration generated from object of interest in the first frame) and γ_a is normalizing factor.

The combined observation likelihood is the product of shape and appearance observation likelihoods as

$$p(I_t|X_t) = p_s(\Gamma_t|X_t)p_a(I_t|X_t). \quad (7)$$

4. Tracking and Change Detection

4.1. Tracking using Auxiliary Particle Filters

In this paper, we use an auxiliary particle filter (APF) for tracking. The APF with sampling/importance resampling was introduced by Pitt and Shephard [16] as a variant of the standard particle filter to improve sample efficiency. An index K for each particle (called the branch index [7]) is introduced as an auxiliary variable. The sampling is performed in two steps. First draw branch index K according to the marginal auxiliary variable weights $\hat{w}_k = Pr(K = k|I_{1:t})$. Then draw X_t from $p(X_t|X_{t-1}^{(k)})$, the prior density based on the branch indices k. The joint sample set is $\{X_t^{(j)}, K^{(j)}\}$ and the corresponding weights are updated by $w_t^{(j)} \propto w_{t-1}^{K^{(j)}} \frac{Pr(I_t|X_t)Pr(X_t|X_{t-1}^{K^{(j)}})}{Pr(X_t,K^{(j)}|I_t)}$. These are then used to get the filtering (posterior) distribution, $Pr(X_t, K = k|I_{1:t})$.

4.2. Detecting Transition Points Between Activities

Activities will change in the course of a long video sequence. The activity changes will cause the PF, with a large enough number of particles, 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 particle filter operates, the new observations would appear to have very large observation noise. Thus the tracking error will increase when the activity changes and this can be used to detect the change times. The tracking error (TE) [1] or prediction error is the distance between the current observation and its prediction based on past observations. In our observation model, TE of tracked configuration $\tilde{Y}_t$ is calculated by

$$TE(\tilde{Y}_t, I_t) = TE_{color}(\tilde{Y}_t, I_t) \circ TE_{shape}(\tilde{Y}_t, I_t), \quad (8)$$

$$\text{where} \quad TE_{color}(\tilde{Y}_t, I_t) = \sum_{l=1}^{k} ||B(ch_l, CH_l)||^2$$

$$\text{and} \quad TE_{shape}(\tilde{Y}_t, I_t) = \sum_{l=1}^{k} ||q_l - f(q_l, G_t)||^2.$$

We combine the shape and color spaces multiplicatively (a heuristic choice that can be replaced with others like addition), so that

$$TE(\tilde{Y}_t, I_t) = \sum_{l=1}^{k} ||B(ch_l, CH_l)||^2 \cdot \sum_{l=1}^{k} ||q_l - f(q_l, G_t)||^2. \quad (9)$$

For the case when the switch from one activity to another is a slow one, the PF dose not lose track very quickly (the tracking error increases slowly). The tracking error will take long to detect the change, and then we use the *Expected (negative) Log Likelihood (ELL)* [19] of the tangent coordinates of the shape. It is calculated as

$$ELL = E[-\log p(v_t)] \quad (10)$$

and is approximated by

$$ELL^N = \frac{1}{N} \sum_{i=1}^{N} v_t^{(i)^T} \Sigma_v^{-1} v_t^{(i)} + K, \quad (11)$$

where K is a constant, N is the number of particles, and Σ_v is the covariance matrix of v. A change is detected when ELL significantly exceeds the learned average value for an activity (equal to the effective rank of Σ_v).

4.3. Reinitialization to New Activity Model

Once a change has been detected, the next problem is to determine the correct activity from the class of previously learned activity models (i.e., the activity template). This is known as the problem of *model switching*, and leads to automatic reinitializaion. Denoting the switching time instance as t_s, and the tracked configuration as $\tilde{Y}_{t_s}$ (which has lost tracking), our reinitialization process is performed automatically as follows:

1. Choose the image at a time before t_s, when TE or ELL is low. Let $\tilde{Y}_{ref}$ denote the tracked configuration of this frame. The contour obtained from this configuration $\tilde{Y}_{ref}$ is used as a reference template, $Temp_{ref}$.

2. On the image at time t_s, I_{t_s}, choose a local region around $\tilde{Y}_{t_s}$. Compute the best match of $Temp_{ref}$ in the region, and denote it as $\bar{I}_{t_s} (\subset I_{t_s})$.

3. Match each learned model to $\bar{I}_{t_s}$ using the deformation and scale model of Sec. 2.2. Denote the best matching configuration as $\tilde{Y}^m, m = 1, ..., M$, where M is the number of all candidate models.

4. Switch to the model of activity m, if $TE_{shape}(\tilde{Y}^m, I_{t_s})$ is the minimum over all M activities.

This process can automatically reinitialize in a changing background, when the change in the activity is gradual (which is the case in most videos).

4.4. Tracking and Change Detection Algorithm

We now outline the main steps of the tracking and change detection algorithm. For simplicity, let us assume that there are two activities in the sequence, A_1 and A_2. Note that A_1 and A_2 may represent different portions of the same activity - specifically, for those activities in which a non-stationary dynamical model is needed. For the first frame in A_1, the region of interest is detected based on the application requirements and the corresponding model for the activity is

determined as in Section 4.3. 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 or ELL) are computed for each frame.

Change Detection When the TE or ELL exceeds a threshold for a few consecutive frames, a change is detected.

Model Switching The procedure in Sec. 4.3 is used to automatically reinitialize and go back to track.

5. Experimental Results

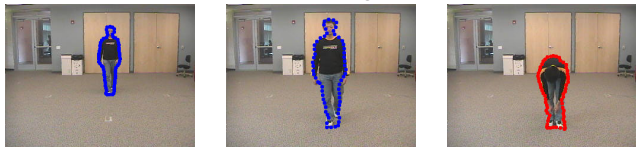
We have performed extensive experiments and analysis of our algorithm. First, we show results on controlled data, and analyze the performance of the proposed method. Next, we show results on different activities in an outdoor environment, including significant occlusions, changing background and multiple interacting people.

5.1. Performance Analysis on Controlled Data

5.1.1 Tracking with Piecewise Stationary Model

We show few examples of PSSA models and the performance of our closed-loop tracking. 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. We extracted the landmarks from this binary image by uniformly sampling on the edge of the silhouette. Once the landmarks are obtained, the shape is extracted using the procedure described in Section 2.1. Using the training data, the parameters of the dynamical models for each activity were learnt using these shape sequences and as explained in Section 2.2. 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.

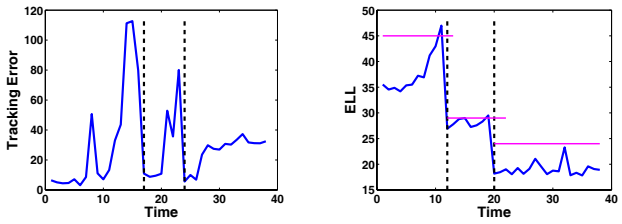
(a) Act1: Bending across



(b) Act2: Walking towards camera and bending down

Figure 3. Tracking results on video data with the PSSA model.

Fig. 3 shows the tracking results for two activities with piecewise stationary models composed of three and two stationary sub-models (i.e., mean shapes), respectively. The number of frames that are used to reinitialize to a new an activity is called the "delay" due to model switching. There is a long delay in the case of slowly changing activities, because the tracking error increases slowly, while for quick changes, the delays are very short. We have explored the use of ELL for slow changes, since it has the potential for reducing the delay. For switching between different subactivities of "bending across", we show the plots for TE and ELL in Fig. 4, where it can be seen that ELL has lesser delay.



(a) (b)

Figure 4. Plot of (a) TE and (b) ELL for Act1. The dotted lines show the switching time instants, and the horizontal lines in (b) show the threshold for ELL. The threshold is set as the rank of Σ_v (see equation 11). ELL detects faster because it is a gradual transition that is tracked by the particle filter.

5.1.2 Comparative Analysis

In order to compare with the open-loop approaches that assume a known transition probability [13, 2, 4, 12, 23], we use our shape-dynamical representation, but incorporate the transition matrix in an open-loop system, instead of the switching module in the feedback path. We discuss the comparison in three cases on a video consisting of several Yoga postures with continuous changes from one posture to another.

• Case 1: Incorrect Transition Probabilities

This is a trivial case where the open-loop approach will lose track. Since the closed-loop approach does not need any such transition probabilities, it will be able to track as long as the new activity model is detected correctly (Sec. 4.3).

• Case 2: Uniform Transition Probabilities

When the accurate transition matrix is unknown, it is reasonable to assume uniform transition probability between activity models. Using the particle filtering framework with a small number of particles, our proposed approach can track through large changes in body posture without knowing the transition probabilities, while an approach like [13] using the same number of particles needs much more time to regain track. Fig. 5 shows the tracking errors with both these approaches. In the open-loop framework, the particles are uniformly distributed within several states. When

the particle number is small (we use 15 particles), the number of particles assigned to each state is few, and the correct state can not dominate. In this case, the open-loop approach needs more time to jump between states before regaining tracking. This can be seen for the second and last postures in Fig. 5 and the tracking results of Fig. 6.

• Case 3: Diagonally Dominant Transition Matrix

In this case, the probability of staying in the same state is higher than switching. This prevents random oscillations between the states. This reintroduces the problem of knowing the probability of staying in the same state as opposed to switching between states. Besides, with a small number of particles, even fewer particles are available for the several possible states when a change happens. This can result in even longer time to regain tracking.

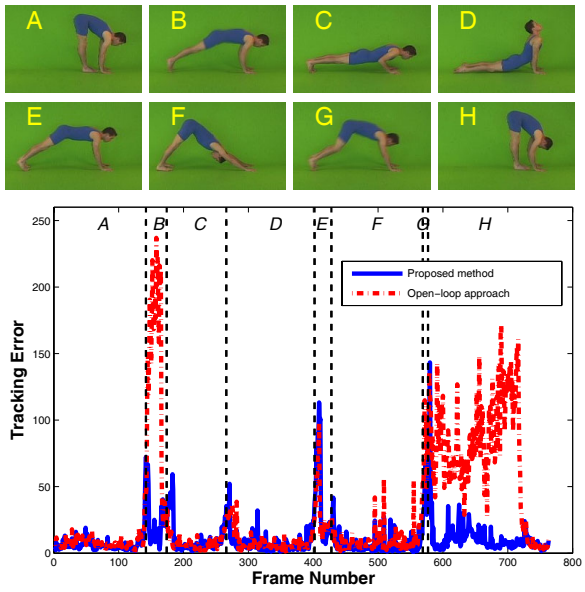
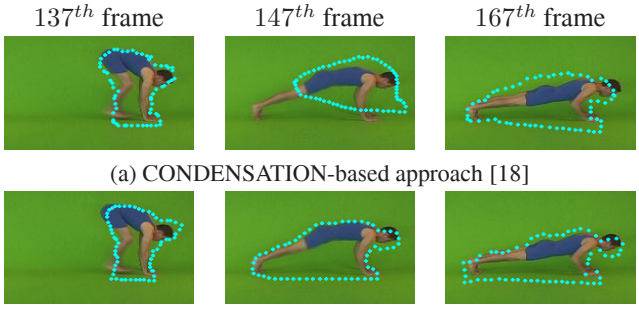


Figure 5. Tracking errors of this Yoga sequence using our proposed approach and an approach like [13] using the same number of particles, and with uniform transition probabilities respectively. It is clearly shown that the latter needs much more time to regain track, where the same dynamical models for describing the activities were used and chosen automatically. The actual switching time instances are shown by black dot lines, and the postures for each time interval are shown on the top. The labels A-H show the different postures.

5.2. Experiments With Outdoor Data

We now show some results with video captured in outdoor environments under different conditions. We have four sequences. The first consists of activities around a car. The second sequence consists of a person walking behind a wall with a package and reappearing without the package. The third is of two people who exchange a box between them in a complex changing background. The fourth consists of multiple people doing different activities simultaneously.



(a) CONDENSATION-based approach [18]

(b) Proposed approach

Figure 6. Tracking results on a video sequence: (a) CONDENSATION-based approach with uniform transition probability, (b) proposed approach which can track without knowing transition probabilities. (a) gets back to track after 30 frames, while our approach is able to maintain tracking continuously.

There are a total of six activities: standing, sitting, walking towards camera, walking parallel to camera, walking with small package in hand, walking with box in hand. Note that scale changes do not affect either the tracking or switching due to the use of shape as a feature.

The first sequence on which we show our results consists of activities of two people: one person walking with a small package in hand, another person first walking towards the camera, and then walking parallel to the camera.

Fig. 7 (a) shows the tracking results on this outdoor sequence, and the switched activity is shown on the top. Top two images are the results for "Person 1", and the three below them are the results for "Person 2". The bottom image of Fig. 7 (a) gives the model switching result for the activities of "Person 2", and the switch from "walking toward camera" to "walking parallel to camera" is clearly shown.

We also did tracking experiments in an outdoor environment with significant occlusions. In Fig. 7 (b), we show a person walking with a box, disappearing behind a wall, and emerging on the other side without the box. We are able to track the person and reinitialize that the box is not present after the person emerges from behind the wall using our model switching approach. The occlusion is detected by the large change in appearance around the observed positions of the persons.

We may face the case that the parts of an object of interest may split and merge (e.g. a person carrying a box, may put it down or give it to someone else). Whenever such kinds of activities are recognized using our model switching, we separate the whole object into parts (e.g. person and box) and track them independently. As an example, for the case of a person carrying a box, the overall observed shape is matched with the shape of a walking person and the two parts separated. Then we decide whether these two parts are together or not based on the relative position and motion between them.

In Fig. 7 (c), we give an example of this case. We

show two persons walking towards each other, one person holding a box and then giving it to the other person. In this sequence, there exist occlusions and significant background changes. The model switching result is shown on the bottom. It is clear that at the beginning Person 1 has a higher probability of holding box, then the probability of Person 2 holding box becomes higher after the exchange. These probabilities are calculated from the tracked results, by computing the distance between the different objects, (i.e. box, Person 1, Person 2) and their relative motion. As an example, if the direction of the motion of box and Person 1 is the same, and their distance is close, then Person 1 is carrying the box.

In Fig. 7 (d), we show a sequence of multiple people doing different activities simultaneously. At the beginning Person 1 is standing and Person 2 is sitting. Then Person 3 and Person 4 pass by, and Person 1 and Person 2 walk toward the camera after that. We can track all the activities along with the changes.

6. Conclusions

In this paper, we proposed a novel framework for tracking of a long multi-activity sequence without knowing the transition probabilities between the activities. Using a dynamical feedback system, we show that our method is able to continuously track sequences of different activities amidst changing background, occlusions and clutter. We use a non-linear, piecewise stationary model defined on the shape of human body contour to represent activities. Tracking is achieved using a composite feature vector comprised on shape, color and motion. Measures are designed to detect changes between activities. We demonstrate the effectiveness of our approach by doing a detailed performance analysis and showing experimental results on real life video of different activities.

References

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

[2] T.-J. Cham and J. M. Rehg. A multiple hypothesis approach to figure tracking. In *IEEE Computer Vision and Pattern Recognition*, 1999.

[3] T. F. Cootes and C. J. Taylor. Active shape models - "smart snakes". In *Proc. British Machine Vision Conf.*, 1992.

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

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

[6] G. D. Finlayson, B. Schiele, and J. L. Crowley. Comprehensive colour image normalization. In *European Conference on Computer Vision*, 1998.

[7] F. Guo and Q. G. Sample-Efficiency-Optimized Auxiliary Particle Filter. In *Proceedings of IEEE Workshop on Statistical Signal Processing*, 2005.

[8] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, Man, Cybernetics - Part C: Applications and Reviews*, 34(3), 2004.

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

[10] J. Jackson, A. Yezzi, and S. Soatto. Tracking deformable moving objects under severe occlusions. In *IEEE Conf. Decision and Control (CDC)*, 2004.

[11] A. Kale, A. Rajagopalan, A. Sundaresan, N. Cuntoor, A. Roy-Chowdhury, A. Krueger, and R. Chellappa. Identification of Humans Using Gait. *IEEE Trans. on Image Processing*, pages 1163–1173, September 2004.

[12] 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.

[13] B. North, A. Blake, M. Isard, and J. Rittscher. Learning and classification of complex dynamics. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(9):1016–1034, 2000.

[14] K. Nummiaro, E. Koller-Meier, and L. V. Gool. Object tracking with an adaptive color-based particle filter. In *Symposium for Pattern Recognition of the DAGM*, 2002.

[15] A. Papoulis. *Probabbility, Random Variables and Stochastic Processes*. McGraw-Hill, Inc., 1991.

[16] M. K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94:590–599, 1999.

[17] Y. Rathi, N. Vaswani, A. Tannenbaum, and A. Yezzi. Particle Filtering for Geometric Active Contours with Application to Tracking Moving and Deforming Objects. In *IEEE Computer Vision and Pattern Recognition*, 2005.

[18] J. Rittscher and A. Blake. Classification of human body motion. In *IEEE Intl. Conf. on Computer Vision*, volume 2, pages 634–639, 1999.

[19] N. Vaswani. Additive change detection in nonlinear systems with unknown change parameters. *IEEE Trans. Sig. Proc.*, March 2007.

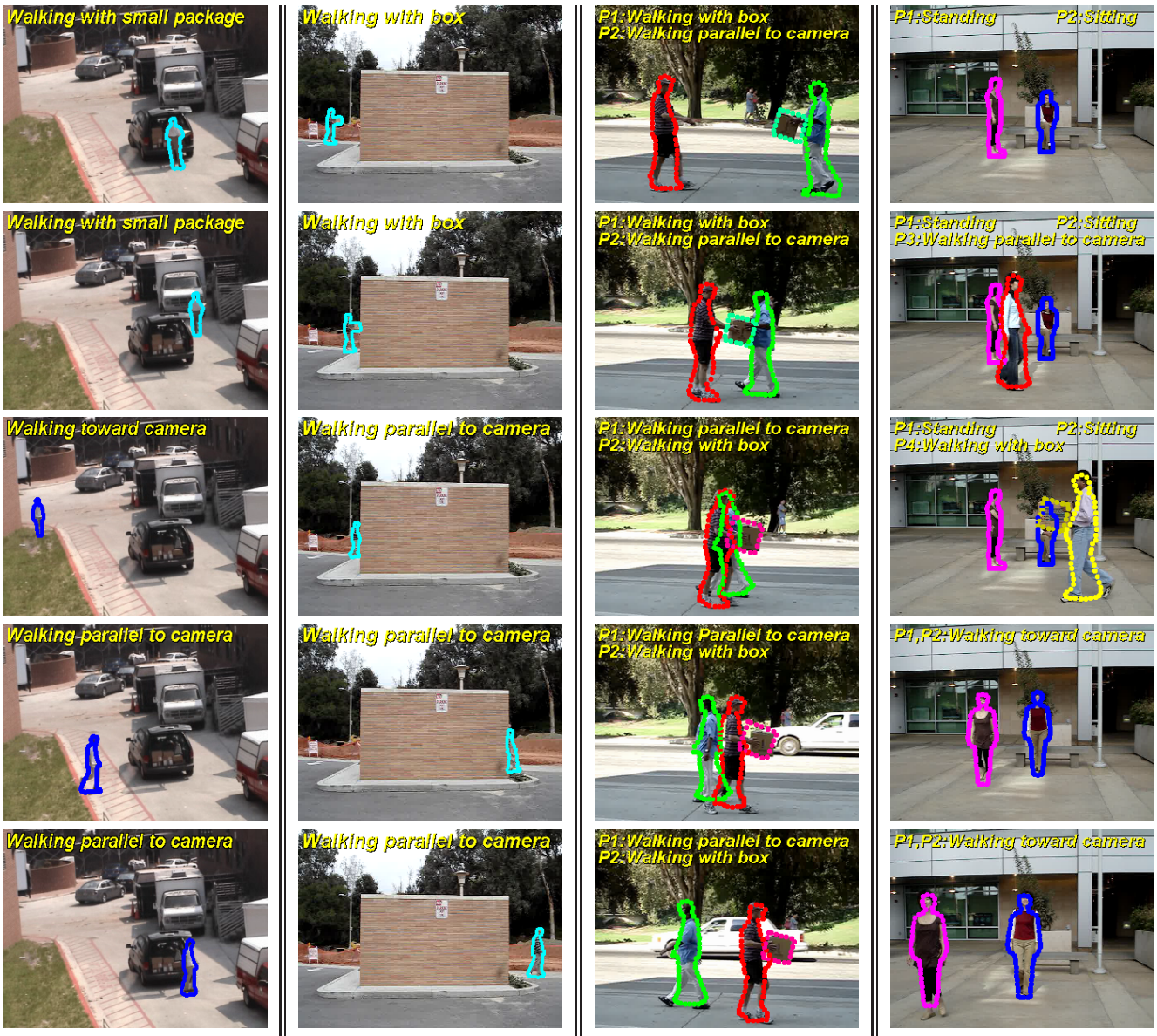
[20] N. Vaswani and R. Chellappa. NonStationary Shape Activities. In *IEEE Conf. Decision and Control*, 2005.

[21] 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*, pages 1603–1616, October 2005.

[22] A. Wilson and A. Bobick. Hidden Markov Models for Modeling and Recognizing Gesture Under Variation. *Intl J. of Pattern Recognition and Artificial Intelligence*, 15(1):123–160, 2000.

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

[24] A. J. Yezzi and S. Soatto. Deformation: Deforming motion, shape average and joint registration and approximation of structure in images. *Intl. Journal of Computer Vision*, 53(2):153–167, 2003.



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

Figure 7. Row 1-5: Tracking results with the activity models used to track shown on the top. Row 6: Model switching results. The labels denote: 1: Standing, 2: Sitting, 3: Walking toward camera, 4: Walking parallel to camera, 5: Walking with small package, 6: Walking with box. (a):Tracking results on an outdoor sequence. (b): A person walking with a box, disappearing behind a wall, and emerging on the other side without the box. We do not track when the person disappears behind the wall. Tracking restarts when we detect that the person has reappeared; (c): Two person walking towards each other, one person holding a box then giving it to the other person. At the first frame, the activity of person 1 is initialized as "walking with a box", then we split box from person and track them separately. (d): Multiple people doing different activities simultaneously. (The figure is best viewed on a monitor.)