COARSE-TO-FINE EVENT MODEL FOR HUMAN ACTIVITIES

Naresh P. Cuntoor and Rama Chellappa

Center for Automation Research
University of Maryland College Park
College Park, MD 20742 USA
cuntoor,rama@cfar.umd.edu

ABSTRACT

We analyze coarse-to-fine hierarchical representation of human activities in video sequences. It can be used for efficient video browsing and activity recognition. Activities are modeled using a sequence of instantaneous events. Events in activities can be represented in a coarse-to-fine hierarchy in several ways, i.e., there may not be a unique hierarchical structure. We present five criteria and quantitative measures for evaluating their effectiveness. The criteria are *minimalism*, *stability*, *consistency*, *accessibility* and *applicability*. It is desirable to develop activity models that rank highly on these criteria at all levels of hierarchy. In this paper, activities are represented as sequence of event probablities computed using the hidden Markov model framework.

Two aspects of hierarchies are analyzed: the effect of reduced frame rate on the accuracy of events detected at a finer scale; and the effect of reduced spatial resolution on activity recognition. Experiments using the UCF indoor human action dataset and the TSA airport tarmac surveillance dataset show encouraging results.

Index Terms— Machine vision, Hidden Markov models, Hierarchical systems.

1. INTRODUCTION

Human activities can be modeled at various scales, ranging from a coarse scale that captures aggregate information to a fine scale that reflects detailed characteristics. Activities can be divided into smaller parts such as sub-activities and events that are localized in time and space. The division proceeds till a level of detail that is suitable to the intended application is attained. At the same time, the model has to effectively address limitations of low video quality and computation time. Coarse-to-fine hierarchical structures have been successfully used to model images and objects ([1], [2], [3]).

There are several ways of decomposing activities to construct a hierarchical structure. It may be useful to develop a set of guiding principles to analyze the effectiveness of a hierarchy. In this paper, we describe criteria based on existing work in 3D shape and action modeling [4],[5]. The usefulness of coarse-to-fine event representation is demonstrated for efficient indexing and browsing. This addresses the effect of degrading video quality (reduced frame rate and low resolution) on event detection. Event probability sequence representation introduced in [6] is used in our experiments.

Several approaches have been proposed for activity modeling in recent years. It is impossible to provide a good overview here. A good review is available in [7]. At the outset, it is perhaps worth distinguishing between approaches that construct a coarse to fine activity model using semantics ([8]), which involves manually specifying the hierarchical structure. Since this is tedious and not readily scalable, we develop a statistical approach using hidden Markov models (HMM).

The paper is organized as follows. Section 2 summarizes the computation of event probability sequences. Section 3 describes the five attributes and quantitative measures for evaluating the effectiveness of hierarchical activity representations. Section 4 demonstrates the efficacy of hierarchical structures using the UCF human action dataset and the TSA airport tarmac surveillance dataset.

2. EVENT MODEL

The objective of an event model is perhaps best illustrated with an example. In commonly performed human activities in an office environment, activities can be thought of as a sequence of instantaneous events. Each event signifies an important change in properties of motion trajectory of objects (speed, direction, heading, etc.) that can be semantically interpreted. For instance, the trajectory traced by a hand in picking up an object lying on the desk can be represented by the following sequence of events: *start*, *extend hand*, *make contact with object*, *grab object*, *withdraw hand*. At a coarse scale, only the *grab object* event may be sufficient. The sequences of events can be obtained in several ways. Each activity and associated events can be manually specified. The process is tedious and may not be scalable. Alternatively, a model-based statistical method can be developed for event detection.

In [6], activities are represented as a sequence of instantaneous, probabilistic events. The representation is called an event probability sequence. It is based on the hypothesis that certain stable transitions at the state level of a HMM denote events. A brief overview and its computation is described in this section.

Let $O = \{o_1, o_2, \ldots, o_T\}$ represent the observation sequence of length T. Specifically, O represents the extracted motion trajectory of an object for T frames, where o_t represents the 2-D pixel location at frame t. The observed sequence O is assumed to be generated by a hidden state sequence $Q = \{q_1, q_2, \ldots, q_T\}$, where $q_t \in \{1, 2, \ldots, N\}$ and N is the model order of the HMM $\lambda = (A, B, \Pi)$. The HMM $\lambda = (A, B, \Pi)$ is characterized by $A = [a_{ij}]$, which is the state transition probability matrix of size $N \times N$; B, the output distribution associated with each state; and Π, the intial state probability vector [9]. For every $t = 1, 2, \ldots, T$, the hidden state q_t can take one of N values so that there N^T possible state sequences. The optimal state sequence is one among these N^T state sequences. Most existing HMM-based techniques use the optimal

This work was partially supported by the ARDA/VACE program under the contract 2004H80200000.

state sequence in modeling. It is not necessary for the optimal state sequence however, to reflect events that have a semantic interpretation. So, we explore other state sequences to detect events such that stable transitions at the state level may be interpreted as events. The number of distinct changes at the state level is finite and equal to $N^2 - N$. The probability of change in passing from state i at time t to state j at time $t+1$ (denoted as $i \longrightarrow j$) can be measured using the variable $\xi_t(i,j)$ used in the Baum-Welch algorithm [9]:

$$\xi_t(i,j) \quad = \quad P(q_t = i, q_{t+1} = j|O, \lambda) \tag{1}$$

For detecting stable changes, (1) is modified so that the probability of the following state transition is measured:

$$\underbrace{i \rightarrow i \rightarrow \ldots \rightarrow i}_{p \text{ frames}} \rightarrow \quad \underbrace{j \rightarrow j \rightarrow \ldots \rightarrow j}_{p \text{ frames}} \tag{2}$$

The event probability variable $\eta_t^p(i,j)$ is defined as follows[1] [6]:

$$\begin{aligned} \eta_t^p(i,j) \quad = \quad & P(q_{t-p} = i, q_{t-p+1} = i, \ldots, q_t = i, \\ & q_{t+1} = j, q_{t+2} = j, \ldots, q_{t+p+1} = j|O, \lambda), \end{aligned} \tag{3}$$

for $p = 2, 3, \ldots, P$, where $P \in \mathbb{N}$. p is referred to as the scale parameter. The most likely change among all possible distinct changes in (3) is said to be an event $e_t^p(k,l)$, i.e.,

$$e_t^p(k,l) = \max_{i \neq j} \eta_t^p(i,j) \tag{4}$$

A peak value in $e_t^p(k,l)$ is interpreted as indicating that an event occurred at time t and the p^{th} level. The event variable can be efficiently computed using the forward and backward variables of the Baum Welch algorithm as follows [6]:

$$\begin{aligned} \eta_t^p(i,j) \quad = \quad & \alpha_{t-p}(i) a_{ii}^p b_i(o_{t-p+1}) b_i(o_{t-p+2}) \ldots b_i(o_t) a_{ij} \times \\ & b_j(o_{t+1}) b_j(o_{t+2}) \ldots b_j(o_{t+p+1}) a_{jj}^p \\ & b_j(o_{t+2}) \beta_{t+p+1}(j) / P(O/\lambda) \end{aligned} \tag{5}$$

where $\alpha_t(i) = P(o_1, o_2, \ldots, o_t, q_t = i|\lambda)$ is the forward variable, $\beta_t(j) = P(o_{t+1}, o_{t+2}, \ldots, o_T|q_t = j, \lambda)$ is the backward variable, $a_{ij} = P(q_t = j|q_{t-1} = i)$ is the one-step transition probability from state i to j and $b_i(o_t) = P(o_t|q_t = i)$ is the probability of observing o_t conditioned on the current state i.

The computation of event probability sequences is briefly summarized. A detailed description is available in [6]. Training consists of three steps: pre-processing to extract features such as motion trajectories; estimating the HMM using the Baum-Welch algorithm [9]; and computation of event probability sequences for every motion trajectory using the trained HMM. During testing, the given test trajectory that is previously not used for training is used to compute candidate event probability sequences for every trained HMM. The candidate event probability sequences are compared with those computed during the training phase using dynamic time warping [10] to allow for non-linear time normalization.

3. ANALYSIS

In this section, we describe five criteria that can be used to determine the effectiveness of a coarse-to-fine event representation. The criteria are *minimalism*, *stability*, *consistency*, *accessibility* and *applicability*. They are based on those developed in and [4] and [11] for 3D shape modeling, and extended by [5] to incorporate the effect of covariates in modeling actions. Activities can be viewed as objects in spatio-temporal domain. So many criteria that are relevant in 3D shape modeling can be readily extended to activity modeling. We propose quantitative measures to evaluate the criteria.

3.1. Minimalism

An event representation should use a minimal number of model parameters required for an application. Generally, the number of parameters needed increases from coarse-to-fine scale. In statistical event models, minimalism can be quantified using a suitable information theoretic criterion such as Akaike information criterion (AIC), bayesian information criterion (BIC) or minimum length description (MDL) [12].

3.2. Stability

Events should be robust to small changes that are caused by imperfections in low-level processing techniques and noise. On the other hand, sensitivity may be desirable at fine scales in order to distinguish between activities with subtle differences.

3.3. Consistency

Events should be consistent at every scale, i.e., the number of events and the location of each event should converge in probability to the true value as the number of samples grows. Consistency can be quantified using a statistical test or a coefficient of consistency such as Pearson's test and Cronbach's correlation coefficient. Pearson's test assumes that the underlying data is normally distributed, which need not be valid in real data. So we use the Cronbach's alpha coefficient, which is defined as follows [13]:

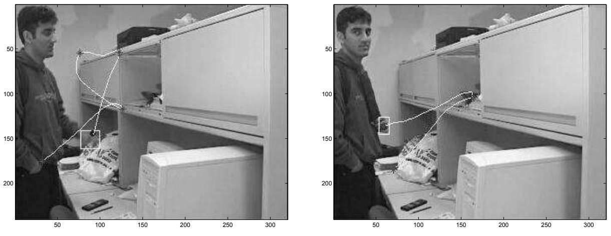
$$\alpha \quad = \quad \frac{K}{K-1}(1 - \frac{\Sigma \sigma_{X_i}^2}{\sigma_Y^2}), \tag{6}$$

where K is the total number of components, $\sigma_{X_i}^2$ is the variance in the i^{th} component and σ_Y^2 is the total variance found by summing across components. Ideally, α should be equal to 1, its maximum value. In the case of event probability sequences, the components are the number of peaks in the event probability and the location of each event. So the number of components need not be constant across activities or across scales in an activity. Each component has to be normalized (mean-subtracted) before computing α.

3.4. Accessibility

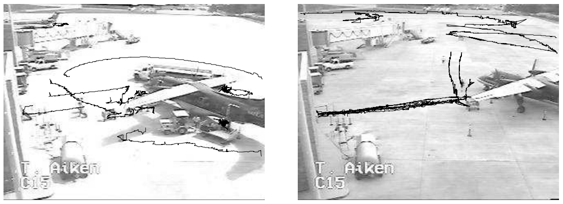
The event representation should be constructed such that fundamental limitations are taken into account. For instance, it may not be possible to determine fine details of an activity because of low video resolution. Also, computationally feasible model estimation algorithms should be available.

[1] There is a minor simplification in notation between the event variable here and that described in [6]. The event variable $\eta_t^p(i,j)$ in [6] corresponds to $\eta_t^{p+1}(i,j)$ here. So, with the new notation, $\eta_t^1(i,j) \overset{def}{=} \xi_t(i,j)$.



(a) (b)

Fig. 1. Sample images from UCF dataset. (a) *open cabinet door*, (b)*pick up object*.



(a) (b)

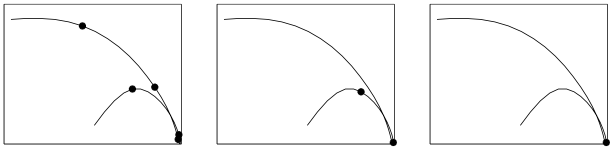
Fig. 2. Sample images from the TSA airport surveillance dataset with automatically extracted motion trajectories. (a) Luggage cart activity around plane and leaves; ground crew walks around plane; and to the gate, (b) Plane enters; luggage cart goes to plane; passengers embark.

3.5. Applicability

Ultimately, application-specific performance measures decide effectiveness of a model. The relative importance of above criteria may be tuned to a specific application. For instance, stability of coarse-to-fine structure may be more important for activity recognition rather than anomaly detection. Applicability can be quantified using performance measures that are such as recognition rate, ROC curves and delay time in event detection.

4. EXPERIMENTS

In this section we highlight advantages of coarse-to-fine hierarchical activity representations for quick video browsing. Hierarchical structures readily lend themeselves to efficient browsing. When browsing, activities can be compared at a coarse scale. Subsequent comparison can be restricted to only those activities that are matched at the coarse scale. Another way to reduce computational time is by reducing the frame rate of test video sequences. These are illustrated using the UCF and TSA datasets.



(a) (b) (c)

Fig. 3. Events detected at different scales for trajectory of *picking up* an object. The scale parameter p varies from (a) $p = 3$, (b) $p = 6$ (c) $p = 8$. Original video resolution is used in event detection.

The UCF dataset consists of 60 video sequences of actions commonly performed in an office environment. The actions are classified into the following classes: *open door*, *close door*, *pick up object*, *put down object*, *pick up object and put down elsewhere*, *pour water*, *erase board*. Figure 1 shows sample images from the dataset along with automatically extracted motion trajectories of the hand. The details of extracting motion trajectories (hand detection, tracking using mean-shift tracker and smoothing using anisotropic filtering) are available in [14].

4.1. Consistency in event detection: effect of reduced frame rate

The UCF human action dataset is used to demonstrate consistency in event detection along with the effect of video quality. Several motion trajectories of each activity is used to train a 4-state left-to-right HMM. Event probability sequences are computed for every motion trajectory as described in section 2. Events detected for *picking up an object* is shown in figure 3. The dots represent events detected along the trajectory traced out by the hand while picking up an object lying on the desk. As expected, fewer events are observed at the coarse scale (larger values of p) compared todetection in the fine scale.

Results: Cronbach's alpha co-efficient is used to quantify consistency in the event representation across multiple observations at a particular scale. The results are summarized in table 1. The effect of reduced frame rate on consistency of events detected is also shown. It is observed that the events detected remain reasonably consistent at a reduced frame rate as well. The consistency measure changes by 1.19% when the frame rate is reduced by half. This suggests that computation of event probability sequences can be speeded up by processing frames at a reduced rate without adversely impacting event detection. If the frame rate is further halved, however, there is a precipitous drop in performance. Also, many quantities in the event computation (section 2) become ill-conditioned because of fewer available data points.

4.2. Coarse-to-fine structure for efficient browsing: effect of reduced spatial resolution

The TSA airport tarmac surveillance dataset is captured by a stationary camera looking at an airport tarmac. It is approximately 120 minutes long and is captured at 30 frames per second and 320×240 resolution. Activities include movement of ground crew personnel, vehicles, planes and passengers embarking and disembarking. Figure 2 shows samples images from the dataset along with the extracted motion trajectories. The following sequence of steps are used to extract motion trajectories: background detection to detect moving objects, foreground segmentation. The KLT algorithm [15] is initialized at the location of the detected moving blobs and object trajectories are obtained.

The TSA dataset is used to illustrate the utility of accessibility and applicability for activity recognition. In far-field surveillance scenarios, moving objects may occupy only a few pixels on the image plane. So it may not be possible to reliably extract trajectories of objects that are far from the camera. At the same time, finer details of motion may not be essential for recognizing activities of interest. With this motivation, we set up an experiment to measure the effect of downsampling on activity recognition. Specifically, motion trajectories may not be accessible at coarse resolution. So we use the

Table 1. Consistency of hierarchical event representation in the UCF human action dataset measured using Cronbach's alpha coefficient. The maximum (and best) value of the coefficient is unity. Results for different activities under two temporal resolutions to demonstrate the trade-off between video quality and reliability of event representation. The average value is weighted by the number of activities in each class.

	Original video	Reduced frame rate (by a factor of 2)	% change
Open door	0.82	0.81	1.22
Pick up object	0.94	0.95	1.06
Put down object	0.81	0.76	6.17
Close door	0.65	0.67	3.08
Pick up & put down elsewhere	0.68	0.70	2.94
Pour water	0.77	0.71	7.79
Erase board	0.91	0.86	5.49
Weighted average	0.84	0.83	1.19

number of foreground pixels as a feature instead of attempting to accurately localize the object.

Activities were classified into four classes of movement: aircraft, passengers, luggage cart and ground crew. HMMs for each class were trained using multiple trajectories at the fine scale, and event probability sequences computed. A 4-state left to right HMM is used. At a coarse scale (spatial resolution is halved), the observed number of foreground pixels are computed and used to train an HMM and compute event probability sequences as before. Using a departure scenario, several synthetic cases were generated and used to train an HMM. This was necessary the dataset consists of only three departure scenarios that occur in entireity.

The testing phase consists of the following steps. Events are compared at the coarse scale by matching event probability sequences using dynamic time warping. If a match is found, event probability sequences are computed at high resolution (original video resolution) using motion trajectories. Detected activities are classified into one of the four classes of movement (defined in the training phase) at the fine scale.

Results: Recognition rates for the aircraft departure activity are presented below. The dataset contains three such scenarios (one of whose trajectories were perturbed and used for training).

(i) At the coarse level, four segments were detected as aircraft departure. The three expected segments were correctly identified. The fourth segment was a false alarm that contained aircraft arrival.

(ii) The segments detected at the coarse level were analyzed further to identify detailed activities. Recognition rates obtained for different activities were as follows. Movement of ground crew: 92%; movement of luggage cart to plane: 100%; passengers embarking: 78%; plane departure: 100%. Some of the passengers were not correctly identified because of truncated trajectories after loss of tracking. In the fourth segment (false alarm), movement of passengers and aircraft were not detected so that matching at the coarse level can be rejected.

5. CONCLUSION

We presented five criteria and quantitative measures for evaluating the effectiveness of coarse-to-fine hierarchical representations of human activities. Experiments using both indoor (UCF human action dataset) and outdoor (TSA airport tarmac surveillance dataset) sequences demonstrate the usefulness of hierarchical structures for activity modeling. It was shown that a reduced frame rate of computing event probability sequences did not have a significant impact on consistency in events detected in the UCF dataset. The effect of reduced spatial resolution on activity recognition was demonstrated using the TSA dataset. At low resolution, it was not possible to reliably extract motion trajectories of individual objects. Instead, aggregate information was used to identify activities in video segments. This reduces computational time required, but compromises the level of detail in modeling. Fine details of activities were extracted using motion trajectories extracted at the original video resolution. As part of future work, we will address minimalism and stability of hierarchical activity representations.

6. REFERENCES

[1] A. Kuijper and L. M. J. Florack, "The hierarchical structure of images," *IEEE Trans. Im. Proc.*, vol. 12(9), pp. 1067–1080, 2003.

[2] D. Gavrila, "Multifeature hierarchical template matching using distance transform," *Proc. IEEE ICPR*, 1998.

[3] F. Fleuret and D. Geman, "Coarse to fine face detection," *IJCV*, vol. 41, pp. 85–107, 2001.

[4] O. Faugeras, *Three dimensional computer vision: A geometric viewpoint*, MIT Press, 1993.

[5] V. Parameswaran and R. Chellappa, "View invariants for human action recognition," *Proc. CVPR*, 2003.

[6] N. P. Cuntoor, B. Yegnanarayana, and R. Chellappa, "Interpretation of state sequences in hmm for activity representation," *Proc. ICASSP*, vol. 2, pp. 709–712, 2005.

[7] A. Roy-Chwdhury R. Chellappa and S. Zhou, *Recognition of humans and their activities using video*, Morgan and Claypool, 2005.

[8] V. T. Vu, F. Bremond, and M. Thonnat, "Temporal constraints for video interpretation," in *Proc. European Conference on Artificial Intelligence*, 2002.

[9] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–285, February 1989.

[10] B. H. Juang, "On the hidden markov model and dynamic time warping for speech recognition - a unified view," *Technical Journal*, vol. 63, pp. 1213–1243, 1984.

[11] D. Marr and H. K. Nishihara, "Representation and recognition of the spatial organization of three dimensional shapes," *Proc. Royal Society, London*, vol. B:200, pp. 269–274, 1978.

[12] G.E.P. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: forecasting and control*, Prentice Hall, 1994.

[13] M. J. Allen and W. M. Yen, *Introduction to measurement theory*, Waveland Press, Long Grove, IL, 2002.

[14] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *IJCV*, vol. 50, no. 2, 2003.

[15] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *IJCAI*, pp. 674–679, 1981.