LEARNING ACTION DICTIONARIES FROM VIDEO

Pavan Turaga and Rama Chellappa

Center for Automation Research, University of Maryland, College Park, MD 20742
{pturaga,rama}@umiacs.umd.edu

ABSTRACT

Summarizing the contents of a video containing human activities is an important problem in computer vision and has important applications in automated surveillance systems. Summarizing a video requires one to identify and learn a 'vocabulary' of action-phrases corresponding to specific events and actions occurring in the video. We propose a generative model for dynamic scenes containing human activities as a composition of independent action-phrases - each of which is derived from an underlying vocabulary. Given a long video sequence, we propose a completely unsupervised approach to learn the vocabulary. Once the vocabulary is learnt, a video segment can be decomposed into a collection of phrases for summarization. We then describe methods to learn the correlations between activities and sequentiality of events. We also propose a novel method for building invariances to spatial transforms in the summarization scheme.

Index Terms— Video Summarization, Activity Analysis

1. INTRODUCTION

Human activity analysis in videos has attracted significant attention in recent years. The goal of a typical system is to recognize activities that are performed in the video from a known set of activities. Most of the current approaches to solve this problem have focused on building parametric or non-parametric models for a set of predefined activities. However, in real world applications, one is not provided with an exhaustive set of the events that may occur in a given setting. Moreover, the system needs to be retrained for every new deployment. These limitations have led researchers to look for unsupervised methods for video indexing and mining. Unsupervised approaches to mine events from videos are extremely challenging. The system is not expected to have or be provided with any knowledge of the domain or the type of activities that might occur in it. The system is expected to discover 'interesting' patterns from the observed scenes within these constraints.

Videos in surveillance settings typically contain multiple activities happening at various locations. Hence, the model should also be able to account for the dependencies, if any, among the activities occurring at different spatial locations. We show that an 'independent basis expansion' is a very rich model for dynamic scenes and can be used in several domains. We also show that the model also provides a natural way of representing activities in a long video as verbs and phrases, thus providing a dictionary of 'action-phrases'.

Prior Work: Activity analysis techniques have primarily used statistical pattern recognition techniques such as HMM's by learning models from a training set [1]. Most earlier unsupervised approaches to video summarization dealt with problems such as shot boundary detection and scene classification [2]. Such approaches are found to work well in genres such as news-videos, but not for human activities. Approaches such as [3], [4] attempt to model the problem of learning activity patterns as one of clustering. Such approaches extract dominant clusters from the videos and consider outlier segments as abnormal. A latent semantic model was used in [5] to recognize actions from long video sequence. A method to recognize activities by modeling multiple activities as independent shapes was presented in [6]. Most such approaches do not explicitly deal with modeling simultaneous activities or inferring the correlations between them. When activities occur simultaneously, typical methods involve tracking each moving object and recognizing the behavior of each by matching with stored models or templates. If the moving objects interact with each other, the problem becomes even harder. In our method, we model dynamic scenes containing human activities as a composition of action-phrases from an underlying dictionary. This allows a natural way of expressing simultaneous activities as a conjunction of several action-phrases. It also allows correlations between activities to be easily identified using co-occurrence statistics of the action-phrases. We propose a completely unsupervised technique to learn the dictionary using an independent basis expansion for video segments.

Organization of the paper: Section 2 describes in detail the model for dynamic scenes and the 'independent basis expansion'. Section 3 discusses the details of learning the action-phrases from a long video. Section 4 presents the summarization algorithm for videos. In Section 5 we present a method to achieve invariance to spatial-transforms in the summarization scheme. Section 6 provides experimental results.

2. MODELING EVENTS IN VIDEOS

Videos with human subjects usually contain several simultaneous activities which may be correlated or independent of each other. In a general setting, one or more of the following may hold true – a) Activities occur simultaneously but are not correlated. e.g. A car entering a parking lot and a person entering a building, b) Activities occur simultaneously and are highly correlated. e.g. A plane arriving at a terminal and the corresponding ground-crew activity, or c) Activities exhibit temporal dependencies. e.g. A plane arriving at a terminal followed by passengers getting off the plane. Any model for dynamic scenes should be able to account for each of these. We model a dynamic scene as a composition of several action-phrases. The compositional model not only allows for modeling independent activities as different action-phrases, but also to decompose a complex scene into its constituent activities. We assume that a video segment can be broken down into constituent action-phrases where each action-phrase would correspond to a particular action being executed by a human or group of humans independent of other activities occurring in the video. Let the set $P = \{P_1, P_2, \ldots P_N\}$ denote the set of action-phrases in a given setting. A dynamic scene D, containing

$n \leq N$ simultaneous activities from the above set can then be seen as a union of the individual action-phrases i.e. $D = p_1 \cup p_2 \cup \ldots \cup p_n$, where $p_i \in P$ and all p_i are distinct.

Since, a given video segment is considered to be a composition of several independent action-phrases we model a given segment as a linear combination of independent basis actions. Suppose $I(x, y, t)$ is a segment of video, the basis expansion is given by

$$I(x, y, t) = \sum_i a_i S_i(x, y, t) \tag{1}$$

where, each S_i corresponds to a particular type of activity. For example, in a parking lot setting, S_1 could correspond to 'A car entering the parking lot' and S_2 to 'A man walking on the sidewalk'. A video segment containing both these activities occurring simultaneously can then be interpreted as a linear combination of these basis activities. Several such representations are possible, but we are interested in a representation that captures not just the motion patterns of individually occurring activities but also their mutual dependence or independence. In the absence of higher level domain knowledge, we use statistical independence as the criterion to discover the independently occurring activities. This constraint also facilitates the learning of the model parameters in equation 1.

2.1. Feature Selection

Direct modeling of an entire video segment as a linear combination of independent bases leads to computational challenges. Firstly, the dimensionality of the data would be extremely high. Secondly, using pixel intensities directly is sensitive to noise and illumination variations. Hence, we extract a feature called the *motion trace*. Given a $M \times N \times T$ space-time volume $I(x, y, t)$, first the moving objects in the scene are extracted by background subtraction. Now, we have a binary space-time volume $I_b(x, y, t)$. The motion trace is recursively defined as

$$M^{(t)}(x, y) = max(\rho * M^{(t-1)}(x, y), I_b(x, y, t)) \tag{2}$$

where ρ represents a *forgetting factor*. The role of ρ is to provide a *trace* of the actor's motion. We typically choose ρ between 0.95 and 0.99. A similar feature, called motion history was used in [7] to perform activity recognition.

In figure 1, we show some examples of the motion traces generated by a person walking on the road and the sidewalk. We see that the feature captures both the spatial location and the motion characteristics of the actor.

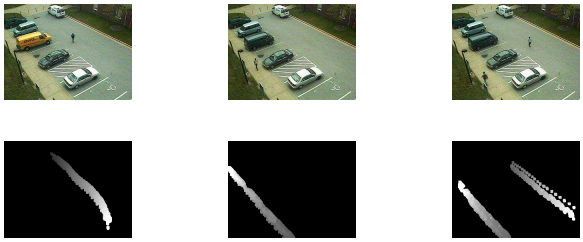

Fig. 1. Top Row: Three frames from video showing a person walking on the road, a person walking on the sidewalk and two persons walking on the sidewalk and the road. Bottom Row: Corresponding Motion Traces

3. LEARNING THE ACTION PHRASES

According to the model in equation 1, each segment is viewed as a mixture of a few underlying independent activities. Given a collection of features (the motion-trace in our case), our goal is to separate the underlying independent bases that generated them i.e. the action phrases. Similar problems have been tackled in the speech and signal processing literature where the problem is commonly termed as Independent Component Analysis (ICA) or Blind Source Separation (BSS). We make use of standard techniques from the ICA literature. To state the problem in general terms, suppose that n linear mixtures $x_1, ..., x_n$ of n independent components are observed –

$$x_j = a_{j1}s_1 + a_{j2}s_2 + ... + a_{jn}s_n, \text{ for all } j. \tag{3}$$

Without loss of generality, it can be assumed that both the mixture variables and the independent components have zero mean. Equation 3 can be written in matrix-vector form as

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{4}$$

where $\mathbf{x}$ is a column vector whose elements are $x_1, ..., x_n$. The matrix $\mathbf{A}$ is called the mixing matrix. ICA involves estimating the mixing matrix $\mathbf{A}$, its corresponding de-mixing matrix $\mathbf{W}$, and the source components $\mathbf{s}$ under the assumption that the sources are statistically independent. We have used the algorithm presented in [8] in our implementations. The learning stage is summarized as follows.

Algorithm 1 Algorithm for Learning the Action Phrases

1: Divide the long video $I_{long}(x, y, t)$ into overlapping segments $\{I_1, I_2, ..\}$ each of length L.
2: For each segment I_i, compute its motion trace M_i.
3: Convert the motion traces into vectors by row ordering.
4: Compute the sources $\{s_i\}$ and the mixing matrix $\mathbf{A}$ using ICA [8].

Note: There exist several methods to estimate the optimal size of the vocabulary i.e. the number of components n. We refer the reader to [9] for more details. Here, we used the AIC criterion to estimate the number of independent components.

4. SUMMARIZING A VIDEO SEGMENT

Each of the learnt independent basis constitutes an action-phrase. The co-ordinates (mixing-ratio) of each vector in the basis provide a description for each segment. These mixing-ratios are also estimated as part of the learning procedure. The co-ordinates in general can be both positive and negative. Since the contribution of a component is indicated by the absolute value of the corresponding co-ordinate, a signature for each segment is created using the absolute values of the co-ordinates and normalizing them by the largest magnitude. A signature for each segment is created by normalizing the vector of mixing-ratios corresponding to that segment.

Now, each segment can be viewed as a document and the basis as a set of phrases. This representation of video segments is similar to the vector-space model (VSM) popularly used in document retrieval literature [10]. The similarity between signatures v_1 and v_2 can be defined as

$$S_{corr}(v_1, v_2) = v_1^T v_2, S_{angle}(v_1, v_2) = cos(\angle(v_1, v_2)) \quad (5)$$

Due to the normalization terms, the cosine similarity measure is more robust to variations in the document size. In our experiments, we use the correlation similarity measure. The results are not significantly different with the cosine measure.

5. INVARIANCE TO SPATIAL TRANSFORMS

In many applications it is desirable to achieve invariance to spatial transforms while learning the action-phrases. First, we prove a result that relates low-level feature transforms to transformations of independent components.

Lemma: Let $\{X(\overline{p})\}$ be a zero-mean random field where $\overline{p} \in D_1 \subseteq R^2$. Let $\{\phi_n^X\}$ be a set of statistically independent basis of X. Let $T : D_2 \longrightarrow D_1$, where $D_2 \subseteq R^2$ be a continuous, one-to-one mapping. Let $\{G(\overline{q})\}, \overline{q} \in D_2$ be a random field derived from X as $G(\overline{q}) = X(T(\overline{q}))$. Then, the set $\phi_n^G(\overline{q}) = \phi_n^X(T(\overline{q}))$ forms a statistically independent basis for G.

Proof: Suppose $\{\phi_n^X\}$ is a set of statistically independent basis of X. i.e.

$$X(\overline{p}) = \sum_i c_i \phi_i^X(\overline{p}) \quad (6)$$

$$p(\phi_i^X(\overline{p}), \phi_j^X(\overline{p})) = p(\phi_i^X(\overline{p}))p(\phi_j^X(\overline{p})), i \neq j \quad (7)$$

Now, suppose $G(\overline{q}) = X(T(\overline{q})), \overline{p} = T(\overline{q})$. Then,

$$G(\overline{q}) = X(T(\overline{q})) = \sum_i c_i \phi_i^X(T(\overline{q})) = \sum_i c_i \phi_i^G(\overline{q}) \quad (8)$$

$$p(\phi_i^G(\overline{q}), \phi_j^G(\overline{q})) = p(\phi_i^X(T(\overline{q})), \phi_j^X(T(\overline{q}))) \quad (9)$$

$$= p(\phi_i^X(T(\overline{q})))p(\phi_j^X(T(\overline{q}))) = p(\phi_i^G(\overline{q}))p(\phi_j^G(\overline{q})), i \neq j \quad (10)$$

Thus, $\phi_n^G(\overline{q}) = \{\phi_n^X(T(\overline{q})\}$ form a statistically independent basis for G.

Application to invariances: Using this result, we can attach a notion of quasi view-invariant similarity between different action-phrases. Let T denote the group of transforms under consideration e.g. affine, homography etc. Let V_1 and V_2 denote two action-phrases (basis vectors). Motivated by the above result, the similarity between two vectors can be expressed as

$$S_{semantic}(V_1, V_2) = \max_T |cos\angle V_1, T(V_2)| \quad (11)$$

where $T(V_2)$ is the transformed version of V_2.

Since, we are interested in learning action-phrases from a single video stream the affine group is sufficient to account for most variations in spatial locations of activities. To perform the optimization, we use Nelder-Mead's simplex method. It is a direct search method and is used when computing gradients is difficult. A good initialization for the optimization problem can be obtained from featureless image registration techniques.

From the learnt set of N action-phrases, we compute the semantic similarity between them as above and arrange them in an $N \times N$ similarity matrix S. We can now define the similarity between individual segments as

$$Sim(v_1, v_2) = v_1^T S v_2 \quad (12)$$

where v_1, v_2 are signatures corresponding to two segments.

6. EXPERIMENTS

In this section, we present the results of experiments based on the proposed approach. The setting of the first experiment consists of a subject executing a series of hand gestures in a near-field setting at varying rates. The subject used four basic hand gestures – {*Raise left hand, Raise right hand, Wave left hand, Wave right hand*} referred to as $\{A, B, C, D\}$ henceforth, and actions done simultaneously such as raise both hands etc. giving rise to eight different activities. The video was collected at about 10 frames per second at a resolution of 720×480.

The set of basis vectors obtained from the proposed algorithm is shown in figure 4. To visualize the learnt activity structures, each segment was embedded into a $2-D$ space using PCA over the learnt mixing weights. Figure 3 shows the resulting embedding. It can be seen that each of the four basic activities is clustered separately. Segments containing simultaneous activities are placed approximately equidistant from the two clusters from which they were formed. This indicates that the model is able to represent simultaneous activities as a combination of the basic ones.

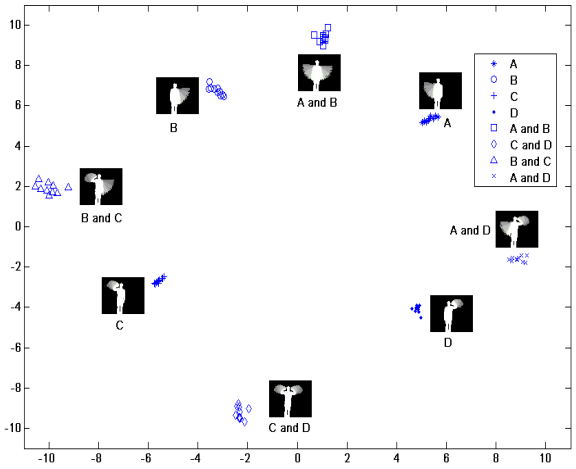

Fig. 3. Clusters of activities

Fig. 4. Learnt basis: Action phrases

Summarization: In figure 5, we show the full summarization of a long video sequence in which the actor performs the above eight activities at random. Each of the four learnt action-phrases is represented with a different color. Occurrence of an action-phrase in a segment is represented by the presence of the corresponding color. Matching of colors indicates that the extracted summary matches well with the ground-truth obtained using manual labeling.

Correlations: We build co-occurrence statistics between the action-phrases from the extracted summary. From the co-occurrence relations, we can identify action-phrases that co-occur and are potentially part of a more complex activity. It was observed that the pairs A-B, A-D, B-C and C-D showed high correlation and hence are potentially parts of a more complex activity. In fact, these pairs

Fig. 2. Detected Unusual Activities. Left: Two persons walking around a vehicle. Right: Person criss-crossing

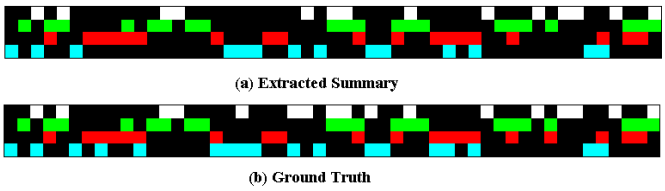

Fig. 5. Results of summarizing a long video sequence. Matching of colors indicates that the right action-phrases have been associated to each video segment. (Figure best viewed in color)

correspond to - Raise both hands, Raise left and wave right, Raise Right and wave left and Wave both hands respectively.

Video Retrieval: In this experiment, we used the TSA dataset which consists of airport surveillance videos. Typical activities that occur are movement of ground crew, vehicles such as fuel cart, luggage cart etc, arrival and departure of planes and passengers getting on and off planes. In the experiment, we selected a query video segment containing a few people getting on a plane and a luggage cart moving toward a distant plane as shown in figure. The best matching segment that was found is shown in figure 6, which also contains people getting on a plane.

Fig. 6. Left: Query segment – Passengers getting on plane and fuel cart moving at a distance, Right: Retrieved Segment – Passengers getting on the plane.

Spatial Invariance: To illustrate the effectiveness of the invariance result we performed a recognition experiment. The setting is the same as described in the first experiment. We generated synthetic data of a change of view by translating and scaling the features and created a long video sequence by concatenating the segments thus obtained. Action-phrases were learnt and the semantic similarity between them was computed. Then, we matched sequences of one view to stored exemplars from the other view. As expected, not compensating for spatial transforms yielded poor results with average recognition accuracy of 10%. After compensating for spatial transforms we obtained an average recognition accuracy of 93.75% which indicates that the proposed spatial invariance method is effective.

Unusual Activity Detection: In the next experiment, we used a 10 minute video sequence of the entrance of a building. The typical activities that were observed were people walking on the side-walk, people walking on the road, and cars entering and leaving the parking lot. Since, the independent basis provides a generative model, we computed the reconstruction errors for each segment. A reconstruction error above a threshold provides an indication of abnormality. For this dataset, the discovered unusual activities are shown in figure 2.

7. CONCLUSIONS

In this paper, we proposed a vocabulary model for dynamic scenes and presented algorithms for unsupervised learning of the vocabulary from long video sequences. We showed the effectiveness of the approach using both far-field and near-field surveillance videos. The results are promising and show that our technique can be used for unsupervised activity indexing as an initial filter for further processing.

8. REFERENCES

[1] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition.," *CVPR*, 1997.

[2] Y. Rui, Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "A unified framework for video summarization, browsing and retrieval," *MERL Technical Report TR2004-115*, 2004.

[3] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," *CVPR*, 2004.

[4] L. Zelnik-Manor and M. Irani, "Event-based video analysis," *CVPR*, 2001.

[5] J. C. Niebles, H. Wang, and L. Fei Fei, "Unsupervised learning of human action categories using spatial-temporal words," *British Machine Vision Conference*, 2006.

[6] A. K. Roy-Chowdhury and R. Chellappa, "A factorization approach to activity recognition," *CVPR Workshop on Event Mining*, 2003.

[7] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[8] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[9] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[10] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of ACM*, vol. 18, no. 11, pp. 613–620, 1975.