FEATURE SELECTION FOR UNSUPERVISED DISCOVERY OF STATISTICAL TEMPORAL STRUCTURES IN VIDEO

Lexing Xie, Shih-Fu Chang

Department of Electrical Engineering
Columbia Univeristy, New York, NY
{xlx, sfchang}@ee.columbia.edu

Ajay Divakaran, Huifang Sun

Mitsubishi Electric Research Labs
Cambridge, MA
{ajayd, hsun}@merl.com

ABSTRACT

We present algorithms for automatic feature selection for unsupervised structure discovery from video sequences. Feature selection in this scenario is hard because of the absence of class labels to evaluate against, and the temporal correlation among samples that prevents the direct estimation of posterior probabilities of the cluster given the sequence. The overall problem of structure discovery [1] is formulated as simultaneously finding the statistical descriptions of structure and locating segments that matches the descriptions. Under Markov assumptions among events, structures in the video are modelled with hierarchical hidden Markov models, with efficient algorithms to jointly learn the model parameters and the optimal model complexity. Feature selection iterates between a wrapper step that partitions the large feature pool into consistent subsets, and a filter step that eliminate redundancy within these subsets, respectively. The feature subsets are then ranked according to the normalized Bayesian Information criteria, and the learning results from these ranked subsets can be evaluated and interpreted by a human observer. Results on soccer and baseball videos show that the automatically selected feature set coincides with those selected with domain knowledge and intuition, while achieving a correspondence comparable to that of supervised learning against manually labelled ground truth.

1. INTRODUCTION

In this paper, we present algorithms for feature selection in unsupervised discovery of statistical temporal structures from video. We define the structure of a time sequence as the repetitive segments that possess consistent deterministic or stochastic characteristics. Though this definition is general to various domains, here we are mainly concerned with the particular domain of video where *structure* represent the syntactic level composition of the video stream. Automatic detection of structures is an inseparable part of video indexing, since it will not only help locate semantic *events* from low-level *observations*, but also facilitate summarization and navigation of the content.

1.1. The structure discovery problem

Given a set of observations, the problem of identifying structure consists of two parts: finding a description of the structure (a.k.a *the model*), and locating segments that matches the description. There are many successful cases under the supervised learning scenario where these two tasks are performed in separate steps - at the *training* step, the algorithm designers manually identify important structures, collect labelled data for training, and apply supervised learning tools to learn a model to describe the structures; at the *classification* step, segments that matches the description are identified. This methodology works for domain-specific problems at a small scale, yet it cannot be readily extended to large-scale data sets in heterogenous domains, as is the case for many video archives.

In our previous work [1], we proposed a unified framework that uses fully unsupervised statistical techniques for automatically discovering salient structures and simultaneously recognizing such structures in unlabelled data. Under certain dependency assumptions, we model the individual recurring events in a video as HMMs, and the higher-level transitions between these events as another level of Markov chain. This hierarchy of HMMs forms a Hierarchical Hidden Markov Model (HHMM), its hidden state inference and parameter estimation are efficiently learned using the expectation-maximization (EM) algorithm. In addition, Bayesian techniques are employed to learn the model complexity, where the search over model space is done with Reverse-Jump Markov Chain Monte Carlo (RJ-MCMC). It archives even slightly better accuracy in recognizing play/break events from soccer video than its supervised counterpart.

1.2. Feature selection for structure discovery

The computational front end in many real-world scenarios extracts a large pool of observations (i.e. features) from the stream, and at the absence of expert knowledge, identifying a subset of relevant and compact features becomes a bottleneck for improving both the learning quality and computation efficiency. Prior work in feature selection for supervised learning mainly divides into filter and wrapper methods according to whether or not the classifier is in-the-loop [2]. For unsupervised learning on spatial data (i.e. assume samples are independent), Xing et. al. [3] iterated between cluster assignment and filter/wrapper methods for known number of clusters. To the best of our knowledge, no prior work has been reported for our specific problem of interest: feature selection for unsupervised learning on temporally dependent sequences with unknown cluster size.

We use a iterative combination of a filter and a wrapper method for feature selection. The first step wraps information gain criteria around HHMM learning, and discover relevant feature groups that are more consistent to each other within the group than across the group; the second step eliminate redundant features that have an approximate Markov blanket within the group; and the last step evaluates each condensed feature group with a normalized BIC, and rank the resulting models and corresponding feature sets with respect to their *a posteriori* fitness. Note this feature selection scheme is independent of the particular learning algorithm, although it is currently implemented around HHMM.

Evaluation on real video data showed very promising results: on soccer and baseball videos, the resulting small number of clusters has a good correspondence with manually labelled classes comparable to those reported in [1]; the highest-scored feature set includes the most distinctive feature by intuition.

The rest of this paper is organized as follows, section 2 discusses the discovery of video structure using HHMM with model adaptation, section 3 presents our feature selection scheme for unsupervised learning on temporal sequences; section 4 includes the test results on several sports videos; section 5 summarizes the work and discusses open issues.

2. LEARNING HIERARCHICAL HIDDEN MARKOV MODELS

We look at videos as temporally highly correlated streams with stochastic observations in discrete event space. Based on observations on the video sequence [1] we adopt a multi-level Markov assumption where each concept is modelled as an HMM and transitions among concepts as another level of Markov chain. These assumptions leads us to HHMM, for which the model structure, the parameter learning and inference, and the model order identification algorithms are summarized in the rest of this section.

2.1. Hierarchical hidden Markov models

HHMM was first introduced [4] as a natural generalization to HMM with hierarchical control structure. As shown in figure 1A, every higher-level state symbol corresponds to a stream of symbols produced by a lower-level sub-HMM; a transition in the higher-level is invoked only when the lower-level model enters an *exit* state (shaded nodes in figure 1A); observations are only produced by the lowest level states. HHMM is also a specialization of Dynamic Bayesian network (DBN), and Figure 1B shows its equivalent DBN representation. In this representation, the state of the model at time t is completely specified by the hidden states Q_t^d at levels $d = 1, \dots D$ from top to bottom, the observation sequence X_t, and the auxiliary *level-exiting* variables E_t^d. Note E_t^d can be turned on only if all lower levels of $E_T^{d+1:D}$ are on.

It is easy to see that the whole parameter set Θ of a D-level HHMM consists of within-level across-level transition probabilities and emission parameters that specifies the distribution of observations conditioned on the state configuration, In our case, the emission parameters are specified by the means μ and covariances σ as emission distributions are Gaussian. The inference and parameter estimation of an HHMM is done through forward-backward iterations similar to those of the HMM, taking into account additional transition and control constraints. The complexity of the algorithm is $O(T)$. Details of the model structure and the estimation algorithm can be found in [1, 5].

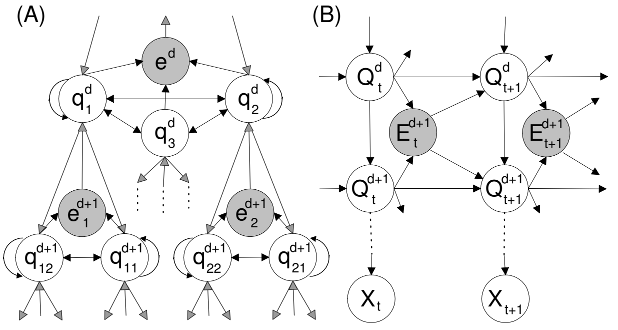

Fig. 1. Graphical HHMM representation at level d and $d+1$ (A)Tree-structured representation; (B)DBN representation, with observations X_t drawn at the bottom. Uppercase letters denote the states as random variables in time t; lowercase letters denote the state-space of HHMM, i.e. values these random variables can take in any time slice. Shaded nodes are auxiliary *exit* nodes that turns on the transition at a higher level.

2.2. Bayesian model selection

In order to automatically determine the complexity of each *event* and the number of *events* in the entire sequence, which maps to the size of the state-space of the HHMM at different levels, we employ Markov chain Monte Carlo(MCMC) algorithm for learning the HHMM structure. In this framework, the optimal size of the HHMM model at all levels is jointly learned with parameter estimation.

MCMC for learning statistical models usually iterates between two steps: (1)The proposal step generates a new structure and a new set of model parameters based on the data and the current model(*Markov chain*) according to certain *proposal distributions* (*Monte Carlo*). In HHMM, the new structure is generated by adding or removing one state in the state hierarchy, or modifying the topological structure while holding the number of states fixed. (2)The decision step computes an acceptance probability α of the proposed new model based on model posterior and proposal strategies, and then this proposal is *accepted* or *rejected* with probability α. MCMC will converge to the global optimum *in probability* if certain constraints are satisfied for the proposal distribution and if the acceptance probability are evaluated accordingly, yet the speed of convergence largely depends on the *goodness* of the proposals. Here we are using a mixture of the EM and MCMC, in place of a full Monte Carlo update of the parameter set and the model size. This brings significant computational savings since EM is more efficient than full MCMC, and the convergence behavior does not seem to suffer in practice. Due to space constraint, the Bayesian model selection algorithm is detailed in [5].

3. FEATURE SELECTION FOR UNSUPERVISED LEARNING

The task of feature selection generally divides into two aspects - eliminating *irrelevant* features and *redundant* ones. Irrelevant features usually disturb the learner and degrade the accuracy, while redundant features add to computational cost without bringing in new information. Furthermore, for unsupervised structure discovery, different subsets of features may relate to different events, and thus the events should be described with separate models rather than being modelled jointly.

3.1. The feature selection algorithm

Denote the feature pool as $F = \{f_1, \dots, f_D\}$, the feature sequence as $X_F = X_F^{1:T}$, and the feature vector at time t as X_F^t. The feature selection algorithm proceeds through the following steps, as illustrated in figure 2:

(1) (Let $i = 1$ to start with) At the i-th round, produce a *reference set* $\tilde{F}_i \subseteq F$ at random, learn HHMM $\tilde{\Theta}_i$ on $\tilde{F}_i$ with model adaptation, perform Viterbi decoding of $X_{\tilde{F}_i}$, and obtain the *reference state-sequence* $\tilde{Q}_i = \tilde{Q}_{\tilde{F}_i}^{1:T}$.

(2) For each feature $f_d \in F \setminus \tilde{F}_i$, learn HHMM Θ_d, get the Viterbi state sequence Q_d, and then compute the information gain (sec. 3.2) of Q_d with respect to the reference se-

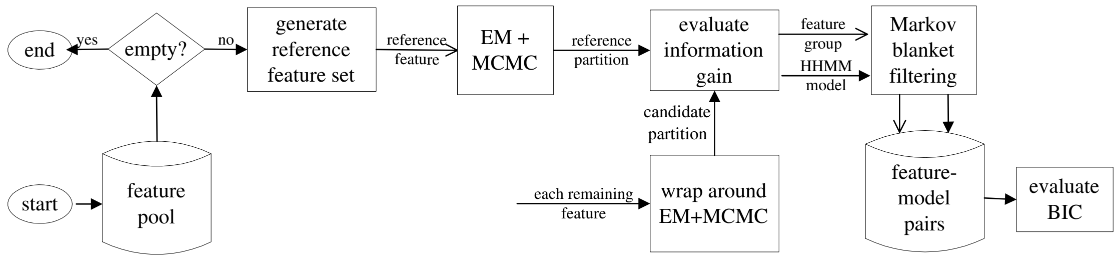

Fig. 2. Feature selection algorithm overview

quence $\tilde{Q}_i$. We then find the subset $\hat{F}_i \subseteq (F \setminus \tilde{F}_i)$ with significant information gain, and form the consistent feature group as the union of the *reference set* and the *relevance set*: $\bar{F}_i \triangleq \tilde{F}_i \cup \hat{F}_i$.

(3) Use Markov blanket filtering in sec. 3.3, eliminate redundant features within the set $\bar{F}_i$ whose Markov blanket exists. We are then left with a relevant and compact feature subset $F_i \subseteq \bar{F}_i$. Re-estimate HHMM Θ_i on X_{F_i}.

(4) Exclude the previous candidate set by setting $F = F \setminus \bar{F}_i$; go back to step 1 with $i = i + 1$ if F is non-empty.

(5) For each feature-model combination $\{F_i, \Theta_i\}_i$, evaluate their *fitness* using the normalized BIC criteria in sec. 3.4, rank the feature subsets and interpret the meanings of the resulting clusters.

After the feature-model combinations are generated automatically, a human operator can look at the structures marked by these models, and then come to a decision on whether a feature-model combination shall be kept based on the meaningful-*ness* of the resulting structures, and the BIC criteria.

3.2. Evaluating information gain

Step 1 in section 3.1 produces a reference labelling of the data sequence induced by the classifier learned over the reference feature set. One suitable measure to quantify the the degree of *agreement* in each feature to the reference labelling, as used in [3], is the mutual information, or the *information gain* achieved by the new partition induced with the candidate features over the reference partition.

$$I(Q_f; \tilde{Q}_i) = H(\tilde{Q}_i) - H(\tilde{Q}_i|Q_f) \quad i = 1, \ldots, N \qquad (1)$$

Here $H(Q)$ is the entropy function over the random variable Q. Intuitively, a larger information gain for candidate feature f suggests that the f-induced partition Q_f is more consistent with the reference partition $\tilde{Q}_i$. After computing the information gain $I(Q_f; \tilde{Q}_i)$ for each remaining feature $f_d \in F \setminus \tilde{F}_i$, we perform hierarchical agglomerative clustering on the information gain vector using a dendrogram [6], look at the top-most link that partitions all the features into two clusters, and pick features that lies in the upper cluster as the set with satisfactory consistency with the reference feature set.

3.3. Finding a Markov Blanket

After wrapping information gain criteria around classifiers built over all feature candidates (step 2 in section 3.1), we are left with a subset of features with consistency yet possible redundancy. The approach for identifying redundant features naturally relates to the conditional dependencies among the features. For this purpose, we need the notion of a Markov blanket[2].

Definition [2] Let F be the set of all features, Q be the class label, f be a feature subset, M_f be a set of random variables that does not contain f, we say M_f is the Markov blanket of f, if f is conditionally independent of all variables in $\{F \cup Q\} \setminus \{M_f \cup f\}$ given M_f.

Computationally, a feature f is redundant if the partition Q of the data set is independent to f given its *Markov Blanket* F_M. In prior work [2, 3], the Markov blanket is identified with the equivalent condition that the posterior probability distribution of the class given the feature set $\{M_f \cup f\}$ should be the same as that conditioned on the Markov blanket M_f only. i.e.,

$$\Delta_f = D(\ P(Q|M_f \cup f) \ || \ P(Q|M_f)\ ) = 0 \qquad (2)$$

where $D(P_1||P_2) = \Sigma_x P_1(x) \log(P_1(x)/P_2(x))$ is the Kullback-Leibler distance between two probability mass functions $P_1(x)$ and $P_2(x)$.

For unsupervised learning over a temporal stream, however, this criteria cannot be readily employed. This is because the posterior distribution of a class depends not only on the current data sample but also on adjacent samples, and conditioning on a neighborhood quickly makes the estimation of the posterior intractable and unreliable. Therefore we use an alternative necessary condition that the optimum state-sequence $Q_{1:T}$ should not change conditioned on observing $M_f \cup f$ or M_f only.

Koller and Sahami have also proved that sequentially removing feature one at a time with its Markov blanket identified will not cause divergence of the resulting set, since if we eliminate feature f and keep its Markov blanket M_f, f remains unnecessary in later stages when more features are eliminated. In practice, there is few if any feature will have a Markov Blanket of limited size, we therefore remove features that induces the least change in the state sequence given the change is small enough. Note computing Markov blanket is a filtering step, since we do not need to retrain the HHMMs for each candidate feature f and its Markov blanket M_f. Given the HHMM trained over the set $f \cup M_f$, the state sequence Q_{M_f} decoded with the observation sequences in M_f only, is compared with the state sequence $Q_{f \cup M_f}$ decoded using the whole observation sequence in $f \cup M_f$. If the difference between Q_{M_f} and $Q_{f \cup M_f}$ is *small enough*(we used 5% in the experiments), then f is removed and M_f is declared to be a Markov blanket of f.

3.4. Normalized BIC

Iterating over section 3.2 and section 3.3 results in feature-model pairs where the disjoint feature subsets are compact and consistent within, and there is one best-effort model fit to each subset. As an extension to using BIC as model posterior in model selection [1], we use a normalized BIC(eq. 3) to measure the *fitness* of the model to the feature subset. Intuitively, the normalized BIC trades off a normalized data likelihood $\tilde{L}$ with model complexity $|\Theta|$. Note $\tilde{L}$ for HHMM is computed in the same forward-backward iterations, normalized with respect to data dimension D, i.e., all the emission probabilities $P(X|Q)$ are replaced with $P'(X|Q) = P(X|Q)^{1/D}$. This normalization is done under the *naive-Bayes* assumption that features are independent given the hidden states. Note the former has weighting factor λ in practice; the latter is modulated by the total number of samples $\log(T)$.

$$\widetilde{BIC} = \tilde{L} \cdot \lambda - \frac{1}{2}|\Theta| \log(T) \tag{3}$$

Note the feature selection scheme is general in that it does not depend on a particular model. The methodology and criteria presented above are model-invariant despite they are tested on HHMM in this paper. Initialization and convergence issues exist in the iterative partitioning of the feature pool. The strategy for producing the random *reference set* $\tilde{F}_i$ in step (1) affects the result of feature partition, as even producing the same $\tilde{F}_i$ in a different sequence may result in different final partitions.

4. EXPERIMENTS AND RESULTS

The feature selection algorithm is tested on soccer and baseball videos. The test videos are all of MPEG-1 format, CIF resolution, 15–32 minutes long. The two soccer clips *Korea* and *Spain* come from MPEG-7 content set, while the baseball clip comes from American TV programs. The initial feature pool is a nine-dimensional feature vector sampled at every 0.1 seconds, including Dominant Color Ratio (DCR), Motion Intensity (MI), the least-square estimates of camera translation (MX, MY), and five audio features - Volume, Spectral roll-off (SR), Low-band energy (LE), High-band energy (HE), and Zero-crossing rate (ZCR).

In this evaluation, we run the unsupervised feature selection and model learning algorithm for each stream. Each resulting feature-model pairs are evaluated and ranked with the modified BIC criteria in section 3.4. Table 1 shows a result snapshot from one sample run. Laid out in each row are the highest-scored feature subset for each clip, the size of the model at the higher level (number of events), and the lower level (the complexity of each event). We also compare the decoded high-level state sequence with a set of manually labelled ground truth, *play* and *break*, defined according to rules of the particular sport. The *correspondence* percentage is evaluated by assigning each of the resulting cluster with its majority ground-truth label, and then compare the resulting sequence with the ground truth.

clip	feature	#events	#children	corr.
Korea	DCR,MX	3	{3,4,7}	75.2%
Spain	DCR,Volume	2	{5,5}	74.8%
Baseball	DCR,MX	2	{6,7}	82.3%

Table 1. A sample run on sports videos

For the two soccer videos, the automatic selected feature set always includes DCR, the most salient feature manually chosen in our prior work [5]. MX approximates the horizontal camera panning motion, which is the most dominant factor contributing to the overall motion intensity (MI) in soccer video. The accuracies are comparable to their counterpart [1] without varying the feature set or even with a fixed feature set and supervised training (75%).

HHMM learning with full model adaptation and feature selection on the baseball video results in three consistent compact feature groups: (a) DCR, MX; (b) Volume, LE; (c) HE, SR, ZCR. It is interesting to see audio features falls into two separate groups, and the visual features are also in a individual group. The BIC score for the *best* group, dominant color ratio and horizontal camera pan, is significantly higher than that of the other two. This selected feature set coincides with our intuition that the status of a baseball game can mainly be inferred from visual information, due to the distinct production syntax of a baseball video. Moreover, the correspondence of the resulting clusters is much higher than that of the soccer videos, suggesting that the production syntax that associates shots to *meaning* does help unsupervised structure discovery.

5. CONCLUSION

In this paper we propose a general feature selection algorithm for unsupervised learning of statistical structure on temporal sequences. The structures in video are modelled with hierarchical hidden Markov models, and the model order at multiple levels are learned with Monte Carlo sampling techniques. We employed an iterative wrapper-filter algorithm that selects the subset of features that is relevant, compact, and consists the best fit to the HHMM model assumptions. We evaluated this algorithm on sports videos, and results are very promising: the clusters matches manually labelled classes, intuitively the most distinctive feature is in the optimal feature set, and evaluation against manually identified structure showed comparable accuracies as its supervised-learning counterpart.

Open issues abound, however: investigating the statistical significance the results, analyzing where the unsupervised learning and the ground truth *mismatch*, and modelling sparse structures are all interesting directions for further investigation.

6. REFERENCES

[1] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, “Unsupervised discovery of multilevel statistical video structures using hierarchical hidden Markov models,” in *Interational Conference on Multimedia and Expo (ICME)*, (Baltimore, MD), July 2003.

[2] D. Koller and M. Sahami, “Toward optimal feature selection,” in *International Conference on Machine Learning*, pp. 284–292, 1996.

[3] E. P. Xing and R. M. Karp, “Cliff: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts,” in *Proceedings of the Ninth International Conference on Intelligence Systems for Molecular Biology (ISMB)*, pp. 1–9, 2001.

[4] S. Fine, Y. Singer, and N. Tishby, “The hierarchical hidden Markov model: Analysis and applications,” *Machine Learning*, vol. 32, no. 1, pp. 41–62, 1998.

[5] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, “Learning hierarchical hidden Markov models for video structure discovery,” Tech. Rep. 2002-006, ADVENT Group, Columbia Univ., http://www.ee.columbia.edu/dvmm/, December 2002.

[6] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.