MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Feature Selection and Order Identification for Unsupervised Discovery of Statistical Temporal Structures in Video

Xie, L.; Chang, S.F.

TR2003-116 September 2003

Abstract

We present algorithms for automatic feature selection and model order identification based on our previous solution to unsupervised structure discovery from video sequences. The overall problem is presented as simultaneously finding the statistical descriptions of structure and locating segments that matches the descriptions. Structures in video was modelled with hierarchical hidden Markov models, and model parameters was efficiently learned using EM. We extend the previous model adaptation scheme to learning not only the complexity of each structure, but also the total number of structures in the stream. Feature selection iterates between a wrapper and a filter method to partition the large feature pool into consistent and compact subsets, where the subsets are then ranked according to a normalized Bayesian Information criteria. Results on soccer videos are very promising: the best feature set agrees with manually identified significant features, the clusters are explainable with respect to manual labels, and the accuracy is comparable with previous works with supervised learning or manually chosen feature sets.

Publication History:

1. First printing, TR-2003-116, September 2003

FEATURE SELECTION AND ORDER IDENTIFICATION 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
Murray Hill, NJ
{ajayd, hsun}@merl.com

ABSTRACT

We present algorithms for automatic feature selection and model order identification based on our previous solution to unsupervised structure discovery from video sequences. The overall problem is presented as simultaneously finding the statistical descriptions of structure and locating segments that matches the descriptions. Structures in video was modelled with hierarchical hidden Markov models, and model parameters was efficiently learned using EM. We extend the previous model adaptation scheme to learning not only the complexity of each structure, but also the total number of structures in the stream. Feature selection iterates between a wrapper and a filter method to partition the large feature pool into consistent and compact subsets, where the subsets are then ranked according to a normalized Bayesian Information criteria. Results on soccer videos are very promising: the best feature set agrees with manually identified significant features, the clusters are explainable with respect to manual labels, and the accuracy is comparable with previous works with supervised learning or manually chosen feature sets.

1. INTRODUCTION

In this paper, we present algorithms for jointly discovering statistical structures, identifying model orders, and finding informative low-level features from video using unsupervised learning. 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, as it will help locate semantic *events* from low-level *observations*. Moreover, further 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 where these two tasks are performed in separate steps. The former is usually referred to as *training*, while the latter, *classification* or *segmentation*. Among various possible models, hidden Markov model (HMM) is a discrete state-space stochastic model with efficient learning algorithms that works well for temporally correlated data streams. HMM has been successfully applied to many different domains such as speech recognition, handwriting recognition, motion analysis, or genome sequence analysis. For video analysis in particular, it has been used for distinguishing TV genres [5], and the high-level play/break structure of soccer games [7].

The structure detection methods above belongs to the category of supervised learning - the algorithm designers manually identify important structures, collect labelled data for training, and apply supervised learning tools to learn the classifiers. 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 [8], we proposed a new paradigm that uses fully unsupervised statistical techniques and aims at automatic discovery of salient structures and simultaneously recognizing such structures in unlabelled data without prior expert knowledge. To the best of our knowledge, it is the first system for unsupervised discovery of temporal statistical structures in video, and it archives comparable, or even slightly better accuracy in recognizing play/break events from soccer video than its supervised counterpart.

In the previous work [8], we presented a unified framework for modelling the temporal dependencies in video, and capturing the generic structure of events. 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).

1.2. Automatic order identification and feature selection

In the HHMM with adaptation scheme presented above, HHMMs were learned over a manually selected set of features with the number of higher-level concepts fixed. Moreover, the number of interesting structures are often unknown *a priori*, then automatically finding the optimal number of high-level concepts is also desirable and will improve the scalability of the learning algorithm to diverse domains. Hence we extend the previous framework to include automatic identification of cluster order in the model adaptation steps, two more types of Monte Carlo moves were choreographed to perform the split and merge of higher level concepts, and the fitness of the new concept space is also evaluated with the Bayesian Information Criteria (BIC).

On the other hand, 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, pick-

ing a subset of relevant and compact features becomes a bottleneck. And automatically identifying informative features, if done, will improve 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 [4]. For unsupervised learning on spatial data (i.e. assume samples are independent), Xing et. al. [9] iterated between cluster assignment and filter/wrapper methods for known number of clusters; Dy and Brodley [2] used scatter separability and maximum likelihood (ML) criteria to evaluate fitness of features. To the best of our knowledge, no prior work has been reported for our problem of interest: unsupervised learning on temporally dependent sequence with unknown cluster size.

We use a combination of filter and wrapper methods for feature selection. The first step is to wrap 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 is to find an approximate Markov blanket for each group, thus eliminating redundant features that does not contribute to uncovering the structure from sequence given its Markov Blanket; and the last step is to evaluate each condensed feature group with a normalized BIC, and rank the resulting models and feature sets with respect to their *a posteriori* fitness.

Evaluation against real video data showed very promising results: on two MPEG-7 soccer videos, the number of clusters that the algorithm converges to is mostly two or three, matching manually labelled classes with comparable accuracies in [8]; the optimal feature set includes the dominant color ratio, the intuitively the most distinctive feature.

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

Videos are temporally highly correlated streams with stochastic observation in discrete concept space [8]. Our attention here is on the subset of *dense* structure, where competing structure elements can be modelled as the same parametric class, and representing their alternation would be sufficient for describing the whole data stream, without needing an explicit *background* model that delineates *sparse* happenings from the majority of the background.

For efficient computation at the cost of minor modelling power, we impose multi-level Markov assumptions [8] where each concept is modelled as an HMM and transitions among concepts as another level of Markov chain. These assumptions leads us to HHMM, where the model structure, parameter learning and inferencing algorithms are summarized in section 2.1, and the model order identification at both levels using MCMC are summarized in section 2.2.

2.1. Hierarchical hidden Markov models

HHMM was first introduced [3] 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. Figure 1B shows the equivalent Dynamic Bayesian Network (DBN) representation of HHMM. 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.

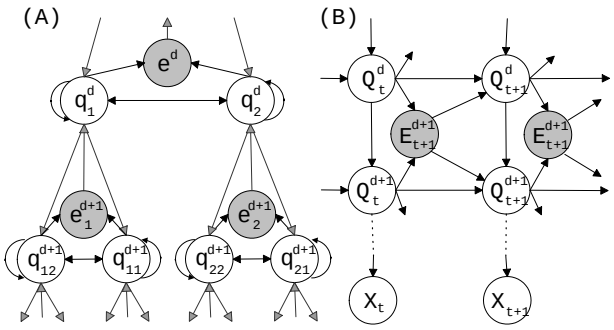

Fig. 1. Graphical HHMM representation at level d and $d+1$ (A)Tree-structured representation; (B)DBN representation, with observations X_t drawn at the bottom. Uppercase letters denote the states as random variables in time t; lowercase letters denote the state-space of HHMM, i.e. values these random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

If the maximum state-space size of any sub-HMM as Q, then the entire configuration of all hierarchical states from the top to the bottom can be represented as a Q-ary D-digit integer k. The whole parameter set Θ of an HHMM then consists of (1) Markov chain parameters Λ^d in level d indexed by the state configuration $k^{(d-1)}$, i.e., transition probabilities A_k^d, prior probabilities π_k^d, and exiting probabilities from the current level e_k^d; (2) emission parameters B that specifies the distribution of observations conditioned on the state configuration, i.e., the means μ_k and covariances σ_k when emission distributions are Gaussian.

The forward-backward algorithm of the HHMM [8, 6] is conducted in a similar manner as those of HMM, essentially operating on the collapsed state space $\{k\}$ taking into account additional transition and control constraints. The state-inference and parameter estimation are then easily built on top of the forward-backward iterations, where the former is a form of multi-level Viterbi algorithm, and the latter involves marginalizing the auxiliary variables within and across difference levels. The complexity of these algorithms is $O(T)$.

2.2. RJ-MCMC for order identification

We employ Markov chain Monte Carlo(MCMC) algorithm in addition to EM for learning HHMMs, in order to address the following problems: (1)EM is known to converge to a local maximum on the likelihood landscape, so it would be stuck in undesirable configurations; (2)The complexity of each concept, i.e. the state-space size of each bottom-level HMM is often unknown empha priori; (3)The number of difference structure elements is unknown either. Yet searching through the entire model space of different orders is intractable. With an MCMC scheme tailored for HHMM, we are able to learn the optimal state-space size of the HHMM model at all levels while learning the parameters.

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*); (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. The two measure spaces under comparison must be aligned if the proposed model is of a different size as the original model, to ensure *Reversibility* of the proposed *Jump* (RJ-MCMC). MCMC will converge to the global optimum *in probability* if certain constraints [1] 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.

Model adaptation for HHMMs is choreographed as follows: (1)Based on the current model Θ, compute a probability profile $P_\theta = [p_{em}, p_{sw}, p_{st}, p_{sb}, p_{mt}, p_{mb}]$, then propose a move among the types {*EM, swap, split-top, split-bottom, merge-top, merge-bottom*} according to the profile P_θ. *EM* is regular parameter update; *Swap* involves swapping the parents of two lower level states associated with different higher-level nodes; *split/merge-bottom* means splitting the emission probability of one of the current bottom level states or merging two of them into one; and *split-top* would randomly partition one higher level state into two and assign its children to either one of the new high-level state, while *merge-top* would collapse two higher-level states into one. (2)Acceptance probability is then evaluated based on model posterior, computed with the Bayesian Information Criteria; for *split* and *merge*, the proposal likelihood and model space alignment also need to be taken into account. Due to space constraint, the EM + RJ-MCMC algorithm is detailed in [6].

Note we are using a mixture of the EM and MCMC, in place of 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.

3. FEATURE SELECTION FOR UNSUPERVISED LEARNING

Feature extraction schemes for audio-visual streams abound, and we are usually left with a large pool of diverse features without knowing which ones are relevant to the concepts in the data. A few features can be selected manually if expert domain knowledge exits, but more often we lack adequate domain knowledge, or the connection between high-level expert knowledge and low-level features are not obvious. Moreover, the task of feature selection is divided into eliminating *irrelevant* features and *redundant* ones, where the former may disturb the classifier and degrade classification accuracy, the latter adds to computational burden without bringing in new information. Furthermore, in the unsupervised structure discovery scenario, different subsets of features may well represent different concepts, and they should be described with separate models rather than modelled jointly.

Hence the scope of our problem, is to select structurally relevant and compact feature subset that fits the HHMM model assumption in unsupervised learning over temporally highly correlated data streams.

3.1. The feature selection algorithm

Denote the feature pool as $F = \{f_1, \ldots, f_D\}$, the data sequence as $\mathbf{X}_F = X_F^{1:T}$, the feature selection algorithm proceeds through these general steps:

(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 $\mathbf{X}_{\tilde{F}_i}$, get the *reference state-sequence* $\tilde{\mathbf{Q}}_i = \tilde{Q}_F^{1:T}$.

(2) For each feature $f_d \in F \setminus \tilde{F}_i$, learn HHMM Θ_d of size $|\tilde{\Theta}_i|$, get the Viterbi state sequence $\mathbf{Q}_d$ compute the information gain (sec. 3.2) of each feature on the $\mathbf{Q}_d$ with respect to the reference partition $\tilde{\mathbf{Q}}_i$. Find the subset $\hat{F}_i \subseteq (F \setminus \tilde{F}_i)$ with significantly large information gain, and keep the union of our *reference set* and the *relevance set* $\bar{F}_i \triangleq \tilde{F}_i \cup \hat{F}_i$ for further processing.

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

(4) Eliminate 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.

3.2. Evaluating information gain

Information gain [9] measures the degree of *agreement* of each feature to the reference partition. We label a partition Q of the original set $\mathbf{X}_F = X_F^{1:T}$ as integers $Q_F^t \in \{1, \ldots, N\}$, let the probability of each part be the empirical portion (eq. 1), and define similarly the conditional probability of the reference partition Q_0 given the partition Q_f induced by a feature f (eq.2). The information gain of feature f with respect to Q^0 is defined as eq. 3.

$$P_Q(i) = \frac{|\{t|q_t = i\}|}{T}; \quad i = 1, \ldots, N \quad (1)$$

$$P_{Q_0|Q_f}(i \mid j) = \frac{|\{t|(q_0^t, q_f^t) = (i, j)\}|}{|\{t|q_f^t = j\}|}; \quad i, j = 1, \ldots, N \quad (2)$$

$$I_g \quad \triangleq \quad H(P_{Q_0}) - \sum_f P_{Q_f} \cdot H(P_{Q_0|Q_f}) \quad (3)$$

Where $H(\cdot)$ is the entropy function. Intuitively, a higher information gain value for feature f suggests that the f-induced partition Q_f is more consistent with the reference partition Q_0.

3.3. Finding a Markov Blanket

After the previous wrapper step, we are left with a subset of features with consistency yet possible redundancy. A feature f is said to be redundant if the partition of the data set is independent to f given its *Markov Blanket* F_M. In prior works [4, 9], Markov blanket is identified with the equivalent condition that the expected KL-divergence between class-posterior probabilities with or without f should be zero.

For unsupervised learning over a temporal stream however, this criteria cannot be readily employed since the temporal correlation prevents us from estimating the posterior distributions by just counting over every feature-label pair. Thus results in two difficulties: (1)The dependency between adjacent observations and class-labels makes the distribution of features and posterior distribution of classes multi-dimensional, and summing over them quickly becomes intractable; (2)We will not have enough data to estimate these high-dimensional distributions. We therefore use an alternative necessary condition that the optimum state-sequence $C_{1:T}$

should not change conditioned on observing $F_M \cup f$ or F_M only. Additionally, as few if any features will have a Markov Blanket of limited size in practice, we sequentially remove features that induces the least change in state sequence given the change is small enough ($< 5\%$).

Note the sequential removal will not cause divergence of the resulting set [4]; and this step is a filtering step since we do not need to retrain the HHMMs for each $F_M \cup f$, Viterbi decoding on only the dimensions of interest would suffice.

3.4. Normalized BIC

Iterating over section 3.2 and section 3.3 results in disjoint small subsets of features $\{F_i\}$ that are compact and consistent with each other. The HHMM models $\{\Theta_i\}$ learned over these subsets are best-effort fits on the features, yet the $\{\Theta_i\}$s may not fit the multi-level Markov assumptions in section 2.

There are two criteria proposed in prior work [2], scatter separability and maximum likelihood (ML). Note the former is not suitable to temporal data since multi-dimensional Euclidean distance does not take into account temporal dependency, and it is non-trivial to define another proper distance measure for temporal data; while the latter is also known [2] to be biased against higher-dimensional feature sets. We use a normalized BIC criteria(eq. 4) as the alternative to ML, which trades off normalized data likelihood $\tilde{L}$ with model complexity $|\Theta|$. Note the former has weighting factor λ in practice; the latter is modulated by the total number of sample values $\log(DT)$; and $\tilde{L}$ for HHMM is computed in the same forward-backward iterations, except all the emission probabilities $P(X|Q)$ are replaced with $P'_{X,Q} = P(X|Q)^{1/D}$, i.e. normalized with respect to data dimension D, under the *naive-Bayes* assumption that features independent given the hidden states.

$$\widetilde{BIC} = \tilde{L} \cdot \lambda - \frac{1}{2}|\Theta| \log(DT) \tag{4}$$

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 different $\tilde{F}_i$ may result in different final partitions. If the dimension of $\tilde{F}_i$ is too low for example, the resulting structure may not be significant and it tends to result in many small feature clusters; on the other hand, if $\tilde{F}_i$ is too large, structures may become too complex, feature subsets maybe too few, and the the result will be hard to interpret.

4. EXPERIMENTS AND RESULTS

The proposed algorithms are tested on two soccer videos taken from MPEG-7 CD, where clip *Korea* is 25 minutes long and *Spain* is 15 minutes. The two semantic events labelled are *play* and *break* [7], defined according to the rules of soccer game. A nine-dimensional feature vector sampled at every 0.1 seconds are taken as the initial feature pool, this include: Dominant Color Ratio (DCR) and 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). We run the feature selection + model learning algorithm on each video stream for five times, with one randomly selected initial *reference feature*. After eliminating degenerate cases such as there are only one feature in the resulting set, we look at the feature-model pair that has the largest *Normalized BIC* value as described in section 3.4.

For clip *Spain*, the selected feature set is {DCR, Volume}, there are two high-level states in the HHMM, each with five lower-level children. Evaluation against the *play/break* labels showed 74.8% accuracy. For clip *Korea*, the selected feature set is {DCR, MX}, with three high-level states and {7, 3, 4} children states respectively. If we assign each of the three clusters the majority ground-truth label it corresponds to (which would be {*play, break, break*} respectively), per-sample accuracy would be 74.5%. This three-cluster results actually matches the previous results [8] with fixed two clusters and manually-selected feature set {DCR, MI}, since the horizontal camera panning contribute to a majority of the whole motion intensity in soccer video, especially when the camera is tracking the ball movement in wide angle. The accuracies are comparable to their previous counterparts [8] without varying the cluster order or the feature set (75%). Yet the small discrepancy may due to:(1) Variability in EM, or the algorithm is yet to converge when maximum iteration is reached; (2)Possible inherent bias may still exist in equation 4 although we are using the same $\lambda = 1/16$ for both algorithms.

5. CONCLUSION

In this paper we propose algorithms for automatic order identification and feature selection in unsupervised learning of statistical structure on temporal sequences. We model the structures in video with hierarchical hidden Markov models, the model order at multiple levels with Monte Carlo sampling techniques. In addition, 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 soccer videos, and results are very promising: the clusters matches manually labelled classes, the 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: The effectiveness of this model applied to other video domains, interpretation of clusters where no pre-defined label is available, and modelling sparse structures are all interesting directions for further investigation.

6. REFERENCES

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, special issue on MCMC for Machine Learning, to appear 2003.

[2] J. G. Dy and C. E. Brodley. Feature subset selection and order identification for unsupervised learning. In *Proc. 17th International Conf. on Machine Learning*, pages 247–254. 2000.

[3] S. Fine, Y. Singer, N. Tishby. The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998.

[4] D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

[5] Y. Wang, Z. Liu, and J. Huang. Multimedia content analysis using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(6):12–36, November 2000.

[6] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Learning hierarchical hidden Markov models for video structure discovery. ADVENT Tech. Report 2002-006, Columbia Univ., http://www.ee.columbia.edu/~xlx/research, December 2002.

[7] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with hidden Markov models. In *Proc. ICASSP*, 2002.

[8] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Unsupervised discovery of multilevel statistical video structures using hierarchical hidden Markov models. Submitted for conference publication, Janurary 2003.

[9] E. P. Xing and R. M. Karp. CLIFF: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. In *Proc. 9-th Conf. Intelligence Systems for Molecular Biology (ISMB)*, pages 1–9, 2001.