UNSUPERVISED DISCOVERY OF MULTILEVEL STATISTICAL VIDEO STRUCTURES USING HIERARCHICAL HIDDEN MARKOV MODELS

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

Structure elements in a time sequence (e.g. video) are repetitive segments with consistent deterministic or stochastic characteristics. While most existing work in detecting structures follow a supervised paradigm, we propose a fully unsupervised statistical solution in this paper. We present a unified approach to structure discovery from long video sequences as simultaneously finding the statistical descriptions of structure and locating segments that matches the descriptions. We model the multilevel statistical structure as hierarchical hidden Markov models, and present efficient algorithms for learning both the parameters and the model structure. When tested on a specific domain, soccer video, the unsupervised learning scheme achieves very promising results: it automatically discovers the statistical descriptions of high-level structures, and at the same time achieves even slightly better accuracy in detecting discovered structures in unlabelled videos than a supervised approach designed with domain knowledge and trained with comparable hidden Markov models.

1. INTRODUCTION

Effective solutions to video indexing require detection and recognition of *structure* and *event* in the video. Where *structure* represents the syntactic level composition of the video content, and *event* represents the occurrences of certain semantic concepts. In this paper, we present a solution to unsupervised structure discovery from video using statistical models. We define the structure of a time sequence as the repetitive segments that possess consistent deterministic or stochastic characteristics. This definition is general to various domains, and structures exist at multiple levels of abstraction. At the lowest level for example, structure can be the frequent triples of symbols in a DNA sequence, or the repeating color schemes in a video; at the mid-level, the seasonal trends in web traffics, or the canonical camera movements in films; and at a higher level, the genetic encoding and controlling regions in DNA sequences, or the game-specific state transitions in sports video. Automatic detection of structures will help locate semantic events from low-level observations, and facilitate summarization and navigation of the content.

1.1. The structure discovery problem

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) [1] 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, different genres in TV programs were distinguished with HMMs trained for each genre [2], and the high-level structure of soccer games (e.g. play versus break) was also delineated with a pool of HMMs trained for each category [3].

The structure detection methods above falls in the conventional 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 diverse domains at a large scale. In this paper, we propose 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. Domain knowledge, if available, can be used to relate semantic meanings to the discovered structures in a post-processing stage. Unsupervised discovery of structures have been applied to gene motif discovery and web stat mining (see reviews in [4]). Only a few instances has been explored for video. Clustering techniques are used on the key frames of shots [5] to discover the story units in a TV drama, yet the temporal dependency of video was not formally modelled. Left-to-right HMMs were stacked into a large HMM in [6, 7] to model temporally evolving recurrent events in ambulatory videos and films, and the resulting clusters correspond to the locations where the video was captured or explosions in film.

1.2. Our approach

In this paper, we model the temporal dependencies in video and the generic structure of events in a unified statistical framework. 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 parameters are efficiently estimated with the expectation-maximization (EM) algorithm. This framework is general in that it is scalable to events of different complexity; yet it is also flexible in that prior domain knowledge can be incorporated in terms of state connectivity, number of levels of Markov chains, and the time scale of the states. In addition, Bayesian learning techniques are used to learn the model complexity automatically, where the search over model space is done with reverse-jump

Markov chain Monte Carlo, and Bayesian Information Criteria is used as model posterior. Evaluation against real video data showed very promising results: the unsupervised approach automatically discovers the high-level structures, along with their statistical descriptions, and at the same time achieves even slightly better accuracy in detecting discovered structures in unlabelled videos than a supervised approach using domain knowledge and HMM models with similar structure.

The rest of this paper is organized as follows: section 2 motivates the use of HHMM from the characteristics of video structures, and discusses the representation and learning of HHMM; section 3 presents the algorithms for Bayesian learning of model structure; section 4 describes the experiments and results on soccer video; section 5 summarizes the work and discusses open issues.

2. MODELLING VIDEO STRUCTURE WITH HHMM

2.1. Characteristics of video structures

Our main attention in this paper is on the particular domain of video, where the structures has the following properties: (1) Video structure is in a discrete state-space, since we humans understand video in terms of concepts, and we assume there exists a small set of concepts in a given domain; (2) The features, i.e., the observations from data are stochastic since segments of video seldom have identical raw features even if they are conceptually similar; (3) The sequence is highly correlated in time, since the videos are sampled in a rate much higher than that of the changes in the scene. In particular, we will focus our attention on the subset of *dense* structure for this paper. By *dense* we refer to the cases 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, i.e., there is no need for an explicit *background* model that delineates *sparse* events from the majority of the background.

Hence, we model stochastic observation in a temporally correlated discrete state space, and use a few weak assumptions to facilitate efficient computation. We assume that states are discrete and Markov within each concept, and observations are associated with states under a fixed parametric form. Such assumptions are justified based on the satisfactory results from several previous work using supervised HMM of similar forms [2, 3]. We also model the transitions of concepts as a Markov chain at a higher level, as this simplification will bring computational convenience at a minor cost of modelling power.

Based on the two-level Markov setup described above, we use two-level hierarchical hidden Markov model to model structures in video. In this model, the higher-level structure elements usually correspond to semantic events, while the lower-level states represents variations that can occur within the same event, and these lower-level states in turn produce the observations, i.e., measurements taken from the raw video, with a Gaussian distribution. Note the HHMM model is a special case of Dynamic Bayesian Networks (DBN), also note the model can be easily extended to more than two levels, and feature distribution is not constrained to mixture-of-Gaussians. In the sections that follow, we will present algorithms that address the inference, parameter learning, and structure learning problems for general D-level HHMMs.

2.2. The structure of HHMM

HHMM was first introduced [8] as a natural generalization to HMM with hierarchical control structure. As shown in Figure 1(A), every higher-level state symbol corresponds to a stream of symbols

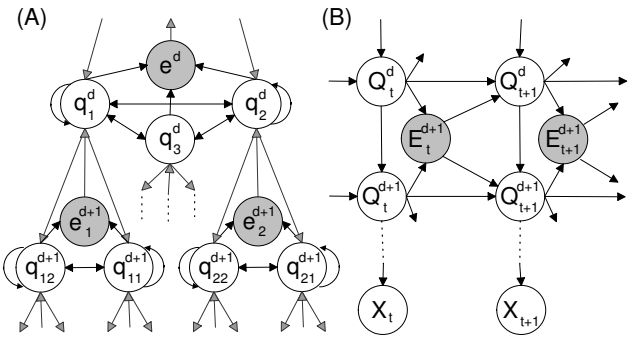

Fig. 1. Graphical HHMM representation at level d and $d+1$ (A)Tree-structured representation; (B)DBN representations, with observations X_t drawn at the bottom. Uppercase letters denote the states as random variables in time t, lowercase letters denote the state-space of HHMM, i.e., values these random variables can take in any time slice. Shaded nodes are auxiliary *exit* nodes that turn on the transition at a higher level - a state at level d is not allowed to change unless the exiting states in the levels below are *on*.

produced by a lower-level sub-HMM; a transition at the high level model is invoked only when the lower-level model enters an *exit* state (shaded nodes in Figure 1(A)); observations are only produced by the lowest level states.

This bottom-up structure is general in the sense that it includes several other hierarchical schemes as special cases. Examples include the stacking of left-right HMMs [6, 7] ; or the discrete counterpart of the jump Markov model with top-down(rather than bottom-up) control structure; HHMM has been applied to solve problems in many domains with different levels of supervision, examples are included in [8, 4].

2.3. HHMM parameter learning and inference

Fine et. al. have also shown that multi-level hidden state inference with HHMM can be done in $O(T^3)$ by looping over all possible lengths of subsequences generated by each Markov model at each level, where T is the sequence length [8]. This algorithm is not optimal, however, an $O(T)$ algorithm has later been shown in [9] with an equivalent DBN representation by unrolling the multi-level states in time (Figure 1(B)). In this DBN representation, the hidden states Q_t^d at each level $d = 1, \ldots D$, the observation sequence X_t, and the auxiliary *level-exiting* variables E_t^d completely specifies the state of the model at time t. The inference scheme used in [9] is the generic junction tree algorithm for DBNs, and the empirical complexity is $O(DT \cdot |Q|^{\lceil 1.5D \rceil} 2^{\lceil 0.5D \rceil})$ where D is the number of levels in the hierarchy, and $|Q|$ is the maximum number of distinct discrete values of any variable $Q_t^d, d = 1, \ldots, D$.

For simplicity, we use a generalized forward-backward algorithm for hidden state inference, and a generalized EM algorithm for parameter estimation based on the forward-backward iterations. This algorithm is derived in a similar fashion as the inference and learning algorithms of HMM; we need to take into account the multi-level transition constraints, and set the parameters space appropriately such that the maximization step in EM has a closed form solutions in each iteration. The parameter set learned by EM includes emission probabilities that associate the observations with the states, Markov chain probabilities at each level, and inter-level transition probabilities. The complexity of this algorithm is

$O(DT \cdot |Q|^{2D})$, with a similar running time as [9] for small D and modest Q. Details can be found in [4].

3. BAYESIAN MODEL ADAPTATION

The EM algorithm for learning HHMM parameters will reach a local maxima of data likelihood, and this algorithm has to operate on a pre-defined model size. Since searching for a global maxima both in the likelihood landscape or across all possible model structures is intractable, we adopt randomized search strategies to address these issues. Markov chain Monte Carlo(MCMC) is a class of such algorithms that has been successfully used to solve high-dimensional optimization problems, especially the problem of Bayesian learning of model structure and model parameters [10]. In this work, we are able to learn the optimal state-space size and the parameters of the entire HHMM with the following MCMC algorithm tailored for HHMMs. The algorithm is outlined in the rest of this section, while the details is in [4] due to space constraint.

MCMC for learning statistical models usually iterates between two steps: (1)The proposal step comes up with a new structure and a new set of model parameters based on the data and the current model(the *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. MCMC will converge to the global optimum *in probability* if certain constraints [10] 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 HHMM involves moves similar to [11] since many changes in the state space involve changing the number of Gaussian kernels that associates states in the lowest level with observations. We included four general types of movement in the state-space, as can be illustrated form the tree-structured representation of the HHMM in figure 1(A): (1)*EM*, regular parameter update without changing the state space size. (2)*Split(d)*, to split a state at level d. This is done by randomly partitioning the direct children (when there are more than one) of a state at level d into two sets, assigning one set to its original parent, the other set to a newly generated parent state at level d; when split happens at the lowest level(i.e. $d = D$), we split the Gaussian kernel of the original observation probabilities by perturbing the mean. (3) *Merge(d)*, to merge two states at level d into one, by collapsing their children into one set and decreasing the number of nodes at level d by one. (4) *Swap(d)*, to swap the parents of two states at level d, whose parent nodes at level $d-1$ was not originally the same. Compared to the RBF network in [11], this special move is needed for HHMM, since its multi-level structure is non-homogeneous within the same size of overall state-space. Moreover, we are not including birth/death moves for simplicity, since these moves can be reached with multiple moves of split/merge.

There are usually three factors in the acceptance probability α, where the first factor directs the moves to the *right* optimal direction, the other two factors ensures the reversibility of the Markov chain moves and the convergence behavior of the sampling algorithm: (1) Model posterior as optimality criteria that evaluates the *fitness* of the new model. Here it is computed with the Bayesian Information Criteria, as in equation (1); (2) A proposal likelihood term that takes into account the model size and the proposal strategies; (3) A model space alignment term introduced by the *Jacobian* of the two spaces, when MCMC moves between two models of different sizes, i.e., *split* and *merge*.

$$BIC = \log(P(X|\Theta)) \cdot \lambda - \frac{1}{2}|\Theta| \log(T) \quad (1)$$

Intuitively, BIC is a trade-off between data likelihood $P(X|\Theta)$ and model complexity $|\Theta| \cdot \log(T)$ with weighting factor λ. Larger models are penalized by the number of free parameters in the model $|\Theta|$; yet the influence of the model penalty decreases as the amount of training data T increases, since $\log(T)$ grows slower than $O(T)$. We empirically choose the weighting factor λ as $1/16$ in the simulations, in order for the change in data likelihood and that in model prior to be numerically comparable over one iteration.

We use a mixture of the EM and MCMC for learning HHMMs, where the model parameters are updated using EM, and the model structure is learned with MCMC. We choose this hybrid algorithm in place of full Monte Carlo update of the parameter set and the model, since MCMC update of parameters will take much longer than EM, and the convergence behavior does not seem to suffer in practice.

4. EXPERIMENTS AND RESULTS

The unsupervised structure discovery algorithm is tested on a 25-minute Korean soccer video clip taken from the MPEG-7 content CD. The two semantic events labelled are *play* and *break* [3], defined according to the rules of soccer game. These two events are dense since they covers the whole time scale of the video, and distinguishing *break* from *play* will be useful for the viewers since *break* takes up about 40% of the screen time. Two manually selected features, dominant color ratio (DCR) and motion intensity (MI) [3], uniformly sampled from the video stream every 0.1 seconds, are used to learn the structure. Note our unsupervised method takes only the raw feature sequence as input, without using the prior knowledge about the existence of play/break concepts or their labels in the data sequence.

4.1. Parameter and structure learning

Here we compare the recognition accuracy of four different learning schemes against the manual labels. (1)Supervised HMM [3]: One HMM per semantic event is trained on manually segmented chunks; and then the video data with unknown boundary are first chopped into 3-second segments, where the data likelihood of each segment is evaluated with each of the trained HMMs; and the final segmentation boundary is obtained after a dynamic programming step taking into account the model likelihoods and the transition likelihoods of the short segments from the segmented data. (2)Supervised HHMM: Individual HMMs at the bottom level of the hierarchy are learned separately, essentially using the models trained in (1); the across-level and top level transition statistics are also trained from segmented data; and then the segmentation is obtained by decoding the Viterbi path from the hierarchical model on the entire video stream. (3)Unsupervised HHMM without model adaptation: An HHMM is initialized with known size of state-space and random parameters; the EM algorithm is used to learn the model parameters; and the segmentation is obtained from the Viterbi path of the final model. (4)Unsupervised HHMM with model adaptation: An HHMM is initialized with arbitrary size of the state-space and random parameters; the EM and RJ-MCMC algorithms are used to learn the size and parameters of the model; state sequence is obtained from the converged model with optimal size. Here we report results separately for (a) model adaptation

in the lowest level of HHMM only, and (b) full model adaptation across different levels as described in section 3.

For algorithms (1)-(3), the model size is set to the optimal model size that (4) converges to, i.e., 6-8

to randomly initialize the HMMs; for unsupervised algorithms(3) and (4), initial bottom-level HMMs are obtained with K-means and Gaussian fitting followed by a grouping algorithm based on temporal proximity. We run each algorithm for 15 times with random start, and compute the per-sample accuracy against manual labels. The median and semi-interquartile range [1] across multiple rounds are listed in table 1.

Learning Scheme	Model type	Super-vised?	Adaptation?		Accuracy	
			Bot.	High	Median	SIQ [1]
(1)	HMM	Y	N	N	75.5%	1.8%
(2)	HHMM	Y	N	N	75.0%	2.0%
(3)	HHMM	N	N	N	75.0%	1.2%
(4a)	HHMM	N	N	Y	75.7%	1.1%
(4b)	HHMM	N	Y	Y	75.2%	1.3%

Table 1. Evaluation of learning schemes (1)-(4) against ground truth using on clip *Korea*

Results showed that the unsupervised learning achieved very close results as the supervised learning case, this is quite promising and surprising since the unsupervised learning of HHMMs is not tuned to the particular ground-truth in the selected video domain. Yet this performance match can be attributed to the carefully selected feature set that well represents the events. For the HHMM with full model adaptation (scheme 4b), the algorithm converges to two to four high-level states, and the evaluation is done by assigning each resulting cluster to the majority ground-truth label it corresponds to. We have observed that the resulting accuracy is still in the same range without knowing how many interesting structures there is to start with. The reason for this performance match lies in the fact that the *additional* high level structures are actually a sub-cluster of *play* or *break*, they are generally of three to five states each, and two sub-clusters correspond to one *larger, true* cluster of play or break.

4.2. Comparing to HHMM with transition constraints

In order to investigate the *expressiveness* of the multi-level model structure, we compare the unsupervised structure discovery performances with that of a similar model with limited transitions. The two model topologies being simulated are (a) The simplified HHMM where each bottom-level sub-HMM is a left-to-right model allowing skips, and cross level transitions can only happen at the first or last node, respectively. (b) The fully connected general 2-level HHMM model as in in figure 1.

Topology (a) is of interest because the transition constraints make the model equivalent to an *collapsed* ordinary HMM, while the general HHMM cannot be collapsed due to the multi-level transition structure. This model simplification leads to an computational complexity of $O(T|Q|^{2D})$, while the general HHMMs will need $O(DT|Q|^{2D})$. Topology (a) also contains the models in [6, 7] as special cases - they used a left-to-right model without skip, and single entry/exit states.

[1] Semi-interquartile as a measure of the spread of at algorithms (4) converges to, i.e. 6-8 bottom-level states per event. For supervised algorithms (1) and the data, is defined as half of the distance between the 75th and 25th percentile, it is more robust to outliers than standard deviation.

The learning algorithm is tested on the same soccer video clip. It performs parameter estimation with a fixed model structure of six states at the bottom level and two states at the top level, over the features set of DCR and MI. Over 5 runs of both algorithms, the average accuracy of the constrained model is 2.3% lower than that of the fully connected model(with average accuracies 71.9% and 74.2%, respectively). This result shows that adopting a fully connected model with multi-level control structures indeed brings in extra modelling power for the chosen domain of soccer videos.

5. CONCLUSION

In this paper we propose algorithms for unsupervised discovery of structure from video sequences. We model the class of dense, stochastic structures in video using hierarchical hidden Markov models, the models parameters and model structure are learned using EM and Monte Carlo sampling techniques. When evaluated on a TV soccer clip against manually labelled ground truth, the fully unsupervised method achieved even slightly better results than its supervised learning counterpart.

Many open issues in stochastic structure discovery using HHMM remains, however: The effectiveness of this model applied to other video domains, incorporation of automatic feature selection techniques, and modelling sparse structures are all interesting directions for further investigation.

6. REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–285, Feb. 1989.

[2] Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Processing Magazine*, vol. 17, pp. 12–36, November 2000.

[3] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden Markov models," in *Proc. Interational Conference on Acoustic, Speech and Signal Processing (ICASSP)*, (Orlando, FL), 2002.

[4] 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.

[5] M. Yeung and B.-L. Yeo, "Time-constrained clustering for segmentation of video into story units," in *International Conference on Pattern Recognition (ICPR)*, (Vienna, Austria), 1996.

[6] B. Clarkson and A. Pentland, "Unsupervised clustering of ambulatory audio and video," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 1999.

[7] M. Naphade and T. Huang, "Discovering recurrent events in video using unsupervised methods," in *Proc. Intl. Conf. Image Processing*, (Rochester, NY), 2002.

[8] 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.

[9] K. Murphy and M. Paskin, "Linear time inference in hierarchical HMMs," in *Proceedings of Neural Information Processing Systems*, (Vancouver, Canada), 2001.

[10] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, pp. 5–43, Jan. - Feb. 2003.

[11] C. Andrieu, N. de Freitas, and A. Doucet, "Robust full bayesian learning for radial basis networks," *Neural Computation*, vol. 13, pp. 2359–2407, 2001.