PATTERN MINING IN VISUAL CONCEPT STREAMS

Lexing Xie[*] *Shih-Fu Chang*

IBM T. J. Watson Research Center Columbia Univeristy

ABSTRACT

Pattern mining algorithms are often much easier applied than quantitatively assessed. In this paper we address the pattern evaluation problem by looking at both the capability of models and the difficulty of target concepts. We use four different data mining models: frequent itemset mining, k-means clustering, hidden Markov model, and hierarchical hidden Markov model to mine 39 concept streams from the a 137-video broadcast news collection from TRECVID-2005. We hypothesize that the discovered patterns can reveal semantics beyond the input space, and thus evaluate the patterns against a much larger concept space containing 192 concepts defined by LSCOM. Results show that HHMM has the best average prediction among all models, however different models seem to excel in different concepts depending on the concept prior and the ontological relationship. Results also show that the majority of the target concepts are better predicted with temporal or combination hypotheses, and there are novel concepts found that are not part of the original lexicon. This paper presents the first effort on temporal pattern mining in the large concept space. There are many promising directions to use concept mining to help construct better concept detectors or to guide the design of multimedia ontology.

1. INTRODUCTION

In this paper we investigate the effective mining and evaluation of pattern mining in large collections of video streams. Important as pattern mining problems are generally regarded, it is much easier to apply data mining algorithms to the collections at hand than to assess the quality and utility of the patterns. i.e., whether the mining results are meaningful and useful. With the recent development of large-scale multimedia concept base and ontology, we can now address the following questions: (1) How useful is algorithm X for mining patterns in video? (2) How well can visual concepts be expressed as combinations of other elementary concepts? (3) How shall we effectively make use of the models to discover useful and novel patterns? Once we have answers to these questions, we will be in a better position to assess the computational feasibility of a multimedia ontology. These results can be used to guide the design of learning algorithms for each concept, and we will also have a rich set of patterns at our disposal for browsing and discovery for novel structures in the data.

Our work closely connects to two research themes: development in the definition and detection of visual concepts, and models for data mining. Finding visual concepts has been an important problem for computer vision and multimedia community. The recent development of defining and detecting the concepts is marked by large-scale annual benchmark [1] and efforts devoted to defining an ontology [2]. Research in multimedia indexing and search has shown that visual concept can improve, among others, search performances [3, 4] or news story tracking [5]. In data mining, plenty of work has been devoted to fast algorithms for finding feature co-occurrences [6], which provide the basis for finding high-confidence causal relationships in the database. However setting a minimum support threshold will exclude the mining of rare and often informative patterns, and the resulting set of co-occurrence rules grows dramatically as the support threshold lowers. Statistical unsupervised models, such as the K-means clustering, mixture model, and dynamic Bayesian networks on the other hand, offers models of controllable complexity that fit the shape of the data automatically.

We investigate effective pattern mining strategies in video streams. We apply different pattern mining models (deterministic and statistic; static and temporal) and devise pattern combination strategies for generating a rich set of pattern hypothesis. We also generate explanations of the statistical patterns in terms of the likelihood of each input feature. We feed the input streams containing 39 visual concept labels [7] to models for unsupervised mining, and evaluate the resulting patterns with an extended ground truth of 192 target concepts [2]. These data come from the Large Scale Concept Ontology for Multimedia (LSCOM) challenge workshop, they provide a valuable basis for assessing the mining strategies in video. In our experiments, more than one-third of the target concepts can be predicted with $F1 > 0.25$ from the elementary observed concepts. Among all 192 target concepts, 86 are best predicted by hierarchical hidden Markov model (HHMM), while 56 and 43 are from frequent itemset mining and K-means, respectively. On average, hierarchical temporal models (hierachical HMM) has showed superior detection performance on a comparable size of rule set while frequent itemset and K-means tend to detect very frequent or very rare concepts well. In addition, HHMM is capable of capturing transition patterns that correspond to the style of video production or story structuring. This paper presents the first effort on temporal pattern mining in the large concept space, where previous work mostly focus on spatial clustering [5], or on low-level feature streams [8].

In the rest of this paper, we will introduce various models for video mining, followed by discussions on the data sets and the experiments.

2. VIDEO MINING

Pattern mining in video concerns with two subproblems: how to abstract from raw video data a set of nominal or numeric attributes, and how to abstract the collection of attributes into perceptually or structurally salient patterns. In this work we focus on the latter, i.e., finding effective unsupervised learning techniques for video. Our mining framework includes an implicit feature selection process by first incorporating all available features and then use unsupervised feature grouping to discover useful feature subsets. There are two categories of mining algorithms: (1) Deterministic algorithms, such as frequent itemsets, seen in traditional data mining scenarios (*FreqItem* for short); (2) Statistical clustering techniques, which in turn include static models such as K-means or mixture models that tries

[*] This material is based upon work funded in part by the U. S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Government.

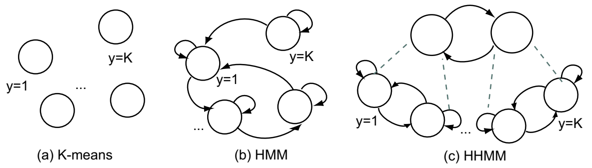

Fig. 1. Models for statistical clustering.

to independently assign cluster labels to each datum, and temporal models such as dynamic Bayesian network that takes into account not only point-wise similarities but also adjacent samples in time.

We now define notations for describing the models and the pattern hypotheses they can induce. We will then introduce four models in the two categories, as well as how enriched pattern hypotheses can be generated by combinations of the clusters.

Let each video v in collection V contain a series of time-stamped shots $\{s_1, \dots, s_T\}$. Assume each shot contains a set of nominal concepts (assume binary for convenience, can be trivially generalized to multi-nominal attributes), such as the presence/absence of *people, face, building, car, etc*. Let this set of concepts be $c = 1, \dots, C$, features in a shot s_t is then written as $x_t = [x_t^1, \dots, x_t^C]'$, where $x_t^c = 0$ means the absence of concepts c in shot t, and $x_t^c = 1$ signals its presence.

2.1. Mining frequent co-occurrences

Frequent itemset mining aims to find item-tuples up to size C that appear in a collection of *transactions* with occurrence frequencies higher than a minimum support threshold $\mathcal{S}_{min}$. It is believed that frequent tuples will provide the basis for finding high-confidence causal relationships among the items. In the video mining case, each shot can be treated as a *transactions*, and this approach can find the most co-occurred concept tuples of any cardinality in the collection.

There are 2^C possible tuples in a database, making it impossible to generate and test all of them even for moderate C. There are, however, topological relationships between the deterministic frequent sets and its subsets so that the search space can be reduced. Closed itemset mining [6] is a fast algorithm for generating non-redundant itemsets (e.g., no superset of these sets have the same support).

We use the IlliMine [9] package to generate the collection of frequent item sets Γ for all concepts in each shot, i.e., $\Gamma = \{\gamma_i \subset C, \forall i\}$. Each itemset γ_i corresponds to a labeling $y_{1:T}(i)$ of the video sequence, where $y_t(i)$=1 if $x_t^c = 1$, $\forall c \in \gamma_i$. In addition to the itemsets in the same shot, we also generate itemsets across adjacent shots by using $[x_t, x_{t-1}]$ as the input to the mining algorithm. The result will be a *Markov* frequent itemset $\Gamma^+ = \{\gamma_i^+ = \gamma_i^+ \cup \gamma_{i,-1}^+,\ \gamma_i^+, \gamma_{i,-1}^+ \subset C, \forall i\}$. The labels $y_t(i)$=1 if $x_t^c = 1$, $\forall c \in \gamma_i^+$ and $x_{t-1}^c = 1, \forall c \in \gamma_{i,-1}^+$.

Frequent itemset mining results in a complete and deterministic list of item tuples seen more than $\mathcal{S}_{min}$ times in the collection. These frequent itemsets can reveal interesting relationships among items. In our video mining results for example, the best predictor for *armed-person* is an itemset containing *military* and *outdoors*, and the best predictor for *highway* is *car* in the last shot along with *outdoors*, *road* and *sky* in the current shot. Albeit being complete, the number of valid tuples often grows drastically as we lower the minimum support threshold, making it hard to identify interesting patterns with modest support.

2.2. Statistical mining from concept streams

The three types of statistical clustering method considered are shown in Fig. 1, where the state-space of these models are numbered 1, …, K. We choose these models for the experiments because (1) they have efficient learning algorithms that scales linearly with the number of data points, (2) they encode an increasingly rich set of data correlations, i.e., from independence to temporal dependence to multi-level temporal dependence. These additional dependencies are useful in improving the detection of concepts, as shown in the examples in the previous subsection.

2.2.1. K-means clustering

K-means clustering maps the set of shots $\{x_1, \dots, x_T\}$ into K clusters by minimizing within cluster variation relative to between-cluster variation. When the input contains nominal concepts, we minimize Hamming distance in concept vector clusters by iteratively computing the distances between each data point and the cluster centroid and assigning points into clusters. The inference in K-means is similar to the expectation-maximization algorithm for the mixture model except that in each iteration the sufficient statistics are taken from hard cluster assignments.

2.2.2. HMM and hierarchical HMM

Since shots in video come as a continuous program stream, we would naturally like to take into account the temporal context by using the adjacent shots to influence the cluster label of the current shot. Hidden Markov model [10] is such a model that encode the dependence between adjacent cluster labels. Hierachical HMM (HHMM) is a generalization of HMM with hierarchical control structure in the hidden states. This two-level dependency in an HHMM allows longer events in video to have within-event variations and transitions.

The size of the state-space in these temporal models represents the number of interesting structures in the data, and it is often desirable to determine the size automatically rather than manually supply a fixed value. We use stochastic search in the state-space to find the optimal model configuration, where the search strategy is generated with reverse-jump Markov chain Monte Carlo (MCMC), and the trade-off criteria between data likelihood and model complexity is based on the minimum description length (MDL) principle. Details of this approach are in our prior work [8].

2.2.3. Concept subset selection

When the number of input features/concepts is large, it is likely that not all of the concepts follow the same clustering distribution or the same evolution dynamics. Hence it is desirable to partition them into mutually consistent subsets that provide different *views* of the same dataset. Extending from the feature selection techniques for supervised learning, and using the mutual information criteria as the similarity measure among the feature dimensions, we cluster the original feature set C into M mutually exclusive feature subsets $C_1, \dots, C_M$, where $\cup_{m=1}^M \{C_m\} = C$. For each feature subset, we learn a HMM/HHMM model with an optimized model structure. A complete description of this approach can be found in [8].

2.2.4. Sequence labeling and hypothesis generation

The clustering algorithm in any of the types above will yield a model φ on one feature subset, with which we can assign each in the original sequence to one of the K mutually exclusive statistical temporal clusters. Furthermore, we can explore the temporal progression of the patterns as well as the different *views* reflected in different feature sets to generate composite hypotheses that may correlate with complex concepts in the domain. The three types of pattern hypotheses are summarized as follows:

1. Original cluster labels. These are the result of applying the cluster model φ to the original feature sequence $x_{1:T}^{1:C}$. i.e., $\varphi(x_{1:T}^{1:C}) = y_{1:T}$, sequence y can be either the cluster labels in K-means, or the maximum-likelihood state sequence

in HMM or HHMM. We can also represent each cluster label with a binary vector $y_t = [y_t(1), \ldots, y_t(K)]'$, $y_t(k) \in \{0, 1\}$, and $\forall k = 1, \ldots, K,\ t = 1, \ldots, T$, $\sum_k y_t(k) = 1$.

2. Temporal composition. We intersect the label sequences across adjacent shots in order to see if the transition statistics in the models indeed capture meaningful progressions in the video sequence. The temporal intersection is defined as follows: $y_{t-1,t}(k_1, k_2) = 1$ if $y_{t-1}(k_1) = 1$ and $y_t(k_2) = 1$.

3. Combining different models. For models φ^{m1}, φ^{m2} learned over two different feature subsets m_1 and m_2, the combination label $y_t^{(m1,m2)}(k_1, k_2) = 1$ if $y_t^{m1}(k_1) = 1$ and $y_t^{m2}(k_2) = 1$. Note that the combination among states in the same model is not computed because they may prefer contradictory descriptions on the same input feature set, e.g., (*people, NOT-crowd*) and (*crowd, parade*).

Let Γ be the set of pattern hypotheses and $|\Gamma|$ be the total number of the resulting pattern hypotheses. $|\Gamma|$ is normally $O(M^2K^2)$ for M feature partitions each with K distinct clusters.

3. THE VIDEO CONCEPT DATASET

The bottleneck for adequate evaluation of pattern mining algorithms is the difficulty in obtaining large amounts of ground truth. The evaluations in this work is made possible by the ARDA sponsored LSCOM workshop [2] for developing an expanded multimedia concept lexicon on the order of 1000.

This large set of concepts are defined and annotated on the *development* part of the TRECVID-2005 [1] video collection. It contains 137 programs (∼80 hours) from three English, two Chinese, and one Arabic channels with a seventeen-day time span from October 30 to November 15, 2005. NIST provided automatically extracted shot boundaries [11] with representative keyframes.

In LSCOM, concepts related to events, objects, locations, people, and programs have been selected following a multi-step process involving input solicitation, expert critiquing, comparison with related ontologies, and performance evaluation. Participants of the process include representatives from intelligence community users, ontology specialists, and researchers on multimedia analytics. In addition, each concept has been qualitatively assessed according to a set of criteria such as utility (usefulness), observability (by humans), and feasibility (by automatic detection).

The annotation efforts in LSCOM is completed in two stages. The first pilot stage named LSCOM-Lite [7] contains 39 concepts obtained by analyzing BBC and TRECVID queries and ensuring that they cover the essential semantic dimensions of news, such as *people, activities, sites, objects, etc.* A collaborative annotation effort was completed in June 2005 to produce concept labels, in the form of the presence or absence of each concept in each key frame, over the TRECVID 2005 development set. Ten of the LSCOM-Lite concepts have been chosen for evaluation in TRECVID 2005 high-level feature detection task [1]. The second stage produces the first version of full LSCOM annotations [2] consisting of 449 unique concepts out of the 834 originally selected, including the 39 LSCOM-Lite concepts. An annotation process similar to that for LSCOM-Lite was completed in late 2005 by student annotators at Columbia University and CMU.

We use the LSCOM-Lite concept labels as the pattern mining input, and a 192-concept subset of LSCOM as our evalulation groundtruth (excluding concepts with no positive examples and those overlapping with LSCOM-Lite). The purpose of this evaluation strategy is to see if patterns discovered from a small set of essential concepts be accurately capture any semantic concept in a much larger space.

4. EXPERIMENTS

We compute each of the four models introduced in Section 2 with different parameter configurations. Figure 2(a) contains a summary of statistics of all the models. The model parameters include the minimum support threshold $\mathcal{S}_{min}$, the cluster size K, and the number of feature partitions M where applicable. Additional settings include the use of concepts in the previous shot in *FreqItem*, denoted with "(-1)", and the use of multiple model sets in the statistical models, denoted with "x3" (see the next subsection). We prune empty and singleton hypotheses (with support zero and one, respectively) from the evaluation, since these degenerate cases will not be reasonable predictors. We record the average number of states $\bar{K}$ (varying due to model selection) and the average number of pattern hypothesis $|\bar{\Gamma}|$ (varying due to model selection and hypotheses pruning) as an indicator of the model complexity and the computational load for evaluating the patterns.

We use 39 LSCOM-Lite concepts as input, 192 LSCOM concepts as targets, and test how capable each clustering technique is in predicting the target concepts. For a pattern sequence $y_{1:T}$ and a groundtruth sequence $z_{1:T}$, $y_t, z_t \in \{0, 1\}$, we use the $F1$ measure derived from precision P and recall R, as its value is often regarded as less biased than precision or recall alone:

$$P = \frac{\sum_t (y_t z_t)}{\sum_t (y_t)}, \quad R = \frac{\sum_t (y_t z_t)}{\sum_t (z_t)}; \quad F1 = \frac{2P \cdot R}{P + R}\ .$$

We take the maximum $F1$ among all candidate patterns in one model (as the majority of the patterns will be a poor predictor for a given target concept), we then average the $F1$ over all the concepts. Except *FreqItem* which generates deterministic results, we simulate five independent runs for each parameter setting in each model and compute the mean and standard deviation of the average F1.

4.1. Predicting concepts and topics

We examine the best predictor for each of the 192 concepts across all runs and all models. HHMM has 86 top-detectors, while HMM, K-means and Freq-item has 12, 56, 43, respectively (as shown in Fig. 2(a), $\#\{$best predictior$\}$ column). We compute the average prior of concepts best predicted by each model (shown in parenthesis). It is obvious that Frequent itemsets tend to perform well in predicting frequent concepts (such as *female/male person, adult*) especially if there are input concepts closely related to the target concept. Such results are quite reasonable – since FreqItem chooses concept groups with high co-occurrence frequency, which in turn naturally has a high likelihood of matching frequent target concepts. In contrast, HHMM and Kmeans are adapt to be capture rare concepts well, such as gymnastics, skiing, jail, all with $F1 > 0.8$.

Averaging the F1 measure across all concepts, we can see that the M HHMM models on partitioned features outperforms the M HMMs and K-means with similar number of patterns. FreqItem has better average performance than any single statistical model, however it is much larger in both the model size and computational load ($|\Gamma|$) than any of the models. In order to make a fair comparison, we simulate a few (say, three) independent runs with different random initializations (denoted by "x3" in Fig. 2(a)) and take the best predictor (max. F1) for each concept across these model instances. We observed that the performance was significantly improved with this multi-initialization strategy, and this multi-HHMM has superior average performance than FreqItem with less than one third of the complexity ($|\Gamma| \sim 4292$ vs. 17540).

4.2. Inspecting the clusters

Among the 192 best predictors, most are either specified by feature subset intersections (81/192) or by temporal transitions (65/192).

model	parameter	$\overline{K}$	$\overline{\|\Gamma\|}$	#{best pred.} (avg. prior)	$\overline{F1}$
FreqItem	S_{min}=500		399	43 (0.077)	0.142
	S_{min}=250	--	817		0.153
	S_{min}=500 (-1)	--	5877		0.147
	S_{min}=250 (-1)	--	17540		0.162
Kmeans	K=20	20	358.2	56 (0.003)	0.102±0.004
	K=50	50	1572.5		0.128±0.002
	K=50 (x3)	150	4732.7		0.159±0.001
HMM	M=1	8.7	74.8	12 (0.029)	0.078±0.007
	M=4	15.4	134.0		0.104±0.006
	M=6	18.5	173.0		0.108±0.006
	M=6 (x3)	55.7	525.3		0.125±0.001
HHMM	M=1	10.6	129.6	81 (0.006)	0.087±0.014
	M=4	30.4	900.8		0.140±0.008
	M=6	36.0	1366.0		0.150±0.016
	M=6 (x3)	111.0	4292.3		**0.181**±0.009

(a)

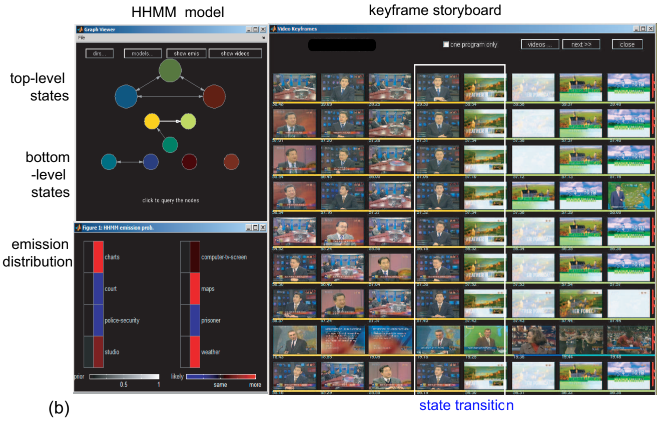

Fig. 2. (a) Performance comparison for different mining schemes. (b) A novel temporal pattern in a visualization system.

Many of these patterns have either confirmed intuitive ontological relationships among concepts, or revealed the production convention of broadcast news. For example, a pair of HHMM states with high confidence in *outdoor, car, road, urban* and the absence of *explosion-fire, natural-disaster, waterscape-waterfront* makes the best *citiscape* predictor; and the best *glacier* detector is from K-means, prescribing *airplane, outdoor, sky* in the current shot, and *natural-disaster, outdoor, snow, truck* and *weather* in the previous shot.

In addition to predicting new concepts, the collection of patterns can be used for browsing the videos and pinpointing novel patterns that are not in the pre-defined concept space. Fig. 2(b) shows such a visualization system we have developed. This is a unique tool for exploring the statistical models such as HMM/HHMM. It not only visualizes the feature distributions, the transitions, but also puts these quantities in the context of the corresponding videos. A research prototype will be made available to public.

The keyframe story board in Fig. 2(b) visualizes the transition pattern from an HHMM state marking the presence of *studio* and *computer-tv-screen* to a state with *weather, maps*, and *charts* (with the two states highlighted in the model panel and frames color-coded with their state labels). Neither of the two states alone are perceptually significant, both of which contain several hundread shots. However the transition between the two states concentrates itself on an interesting production pattern that transits from news anchor to weather report. This suggests that (1) there are salient patterns that can only described by temporal relationships; (2) HHMM or similar temporal models tend to capture consistent but rare temporal patterns, and in some cases, it converges to context-specific production conventions.

In summary, evaluations of temporal pattern mining on a large-scale concept ontology and news topics show that HHMM has superior performance in detecting rare concepts and topic. Partitioning a larger input space into small feature subsets improves concept detection. The HHMM clusters discovered not only reveal interesting ontological relationships among the input and the target concepts, they may also capture structural information not otherwise captured by static concepts or detectable by static models.

5. CONCLUSION

We present an investigation of pattern mining and evaluation in large-scale video concept streams. We have used four different mining algorithms and evaluated on 192 concepts from LSCOM. Results show that HHMM has the best average prediction among all models, however different models seem to excel in different concepts depending on the concept prior and the ontological relationship. Future work may include extending the mining input to actual multimedia observations instead of concept labels, or extending the combination strategies for the models learned.

6. REFERENCES

[1] The National Institute of Standards and Technology (NIST), "TREC video retrieval evaluation," 2001–2005. http://www-nlpir.nist.gov/projects/trecvid/.

[2] "LSCOM lexicon definitions and annotations version 1.0, DTO challenge workshop on large scale concept ontology for multimedia," Tech. Rep. 217-2006-3, Columbia University, March 2006.

[3] S.-F. Chang, W. Hsu, L. Kennedy, L. Xie, A. Yanagawa, E. Zavesky, and D. Zhang, "Columbia University TRECVID-2005 video search and high-level feature extraction," in *NIST TRECVID Workshop*, (Gaithersburg, MD), November 2005.

[4] A. Amir, G. Iyengar, J. Argillander, M. Campbell, A. Haubold, S. Ebadollahi, F. Kang, M. R. Naphade, A. Natsev, J. R. Smith, J. Teic', and T. Volkmer, "Ibm research trecvid-2005 video retrieval system," in *NIST TRECVID Workshop*, (Gaithersburg, MD), November 2005.

[5] J. R. Kender and M. R. Naphade, "Visual concepts for news story tracking: Analyzing and exploiting the nist trecvid video annotation experiment.," in *CVPR (1)*, pp. 1174–1181, 2005.

[6] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data Min. Knowl. Discov.*, vol. 8, no. 1, pp. 53–87, 2004.

[7] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann, "a light scale concept ontology for multimedia understanding for TRECVID 2005," tech. rep., IBM Research, 2005.

[8] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, *Unsupervised Mining of Statistical Temporal Structures in Video*, ch. 10. Kluwer Academic Publishers, 2003.

[9] Data Mining Research Group, University of Illinois at Urbana-Champaign, "Illimine data mining package," December 2005. http://illimine.cs.uiuc.edu/.

[10] 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.

[11] C. Petersohn, "Fraunhofer HHI at TRECVID 2004: Shot boundary detection system," in *NIST TRECVID Workshop*, (Gaithersburg, MD), November 2004.