# Classification of Summarized Videos using Hidden Markov Models on Compressed Chromaticity Signatures

Cheng Lu
School of Computing Science
Simon Fraser University
Vancouver, B.C., CANADA V5A 1S6
(604) 291-4682 Fax (604) 291-3045
clu@cs.sfu.ca

Mark S.Drew
School of Computing Science
Simon Fraser University
Vancouver, B.C., CANADA V5A 1S6
(604) 291-4682 Fax (604) 291-3045
mark@cs.sfu.ca

James Au
School of Computing Science
Simon Fraser University
Vancouver, B.C., CANADA V5A 1S6
(604) 291-4682 Fax (604) 291-3045
Ksau@sfu.ca

## ABSTRACT
Tools for efficiently summarizing and classifying video sequences are indispensable to assist in the synthesis and analysis of digital video. In this paper, we present a method for effective classification of different types of videos that uses the output of a concise video summarization technique that forms a list of keyframes. The summarization is produced by a method recently presented, in which we generate a universal basis on which to project a video frame feature that effectively reduces any video to the same lighting conditions. Each frame is represented by a compressed chromaticity signature. A multi-stage hierarchical clustering method efficiently summarizes any video. Here, we classify TV programs using a trained hidden Markov model, using the keyframe plus temporal features generated in the summaries.

## General Terms
Design.

## Keywords
Video type classification, hidden Markov models, compressed chromaticity signature, temporal feature.

## 1. INTRODUCTION
Video content summarization and classification is a necessary tool for efficient access, understanding and retrieval of videos. Different methods have been proposed in the literature for video program classification into predefined categories such as a commercial detection system [1] and a rule-based basketball video indexing system [2]. However, like most other research work on video classification, advantage was not taken of temporal features in video, a very powerful cue in understanding video content. Therefore we explore the use of hidden Markov models (HMM) for video classification in order to apprehend the temporal information along with the visual information in video.



Previously, we successfully set out a novel illumination-invariant color histogram approach that performs good video characterization [5]. In this method, we form a 12-vector chromaticity signature for any video frame. On the basis of these coefficients we produce keyframe-based succinct summarized expressions for video using a multistage hierarchical clustering algorithm [5]. Here we extend this work to provide the capacity to perform *semantic* content discrimination tasks for video. After video characterization and summarization, we obtain two types of features: (1) chromaticity signatures for keyframes, each of which represents a scene; (2) temporal features including the duration of any scene in a video and *transition characteristics between scenes*. We present a novel method that applies HMM to integrate the two features for video classification. This is motivated by the fact that a certain type of videos usually contains a set of frequent scenes that have similar visual information, such as news and basketball games, and also in most situations those types of videos have their individual stable temporal pattern consisting of scene duration and transition characteristics.

The hidden Markov model is a popular technique widely used in pattern recognition [6]. It has good capability to grasp temporal statistical properties of stochastic processes. The essence of the HMM process is to construct a model that explains the occurrence of observations (symbols) in a time sequence and use it to identify other observation sequences. Some researchers have applied HMM for video analysis and classification. In Nevenka's study [7] using face and text tracking, HMM models can be formed using face and text trajectories and then classify the given video into one of four categories of TV programs: news, commercial, sitcom and soap. The key point of this approach is that the video content for these types of TV programs have to be satisfactorily characterized by capturing face and text trajectories. Boreczky [8] built an HMM framework using audio and image features for video classification. Although the use of both audio and visual features can improve classification accuracy, it can make the system complicated and hard to maintain and extend. Also, because the visual features are extracted for every frame, the HMM process needed to carry a great deal of information about the detailed variance between frames yet lacked consideration of the entire visual trajectory.

In this paper, we set out a video classification method, based on the hidden Markov model, which utilizes the chromaticity signatures of keyframes from summarized video and effectively

grasps the entire temporal feature patterns for different types of videos. First, we use the illumination-invariant color histogram video characterization method proposed in [5] to produce a 12-vector feature for each frame; second we effectively carry out video summarization using a multistage hierarchical clustering, obtaining keyframes. Finally, we perform the video classification task using a hidden Markov model. In our experiment, we apply our method to the task of classifying television programs into the four categories: news reports, commercials, live basketball games, and live football games.

The rest of the report is organized as follows. Section 2 presents our chromaticity signature computation and extraction. Section 3 describes our summarization schema. Video classification based on HMM approach is discussed in section 4. Experimental results are given in section 5 and in section 6 we present the conclusion and an outline of future work.

# 2. ILLUMINATION-INVARIANT VIDEO SUMMARIZATION

We had developed a new low-dimensional video frame feature that is more insensitive to lighting change, motivated by color constancy work in physics-based vision, and apply the feature to keyframe production using hierarchical clustering. The point, vis-à-vis video summarization is that any video is effectively moved into the same lighting environment, making it meaningful to project video features onto a precomputed universal basis set.

Lighting is first discounted by normalization of color-channel bands[4]. This step approximately but effectively removes dependence on both luminance and lighting color. Then image frames are moved into a chromaticity color space. As well as reducing the dimensionality of color to 2 this also has the effect of removing shading. In order to make the method fairly robust to camera and object motion, and displacements, rotations, and scaling, we go over to a 2D histogram derived from DC components of frames. Chromaticity histograms are then compressed –i.e., we treat the histograms as images (see [4]). Here, we use a wavelet-based compression because this tends to strike a balance between simple low-pass filtering and retaining important details. Using a 3-level wavelet compression we arrive at 16*16 histograms.

However, we found that compression of histograms could be improved if the histograms are first binarized, i.e., entries are replaced with 1 or 0. The rationale for this step is that chromaticity histograms are a kind of color signature for an image, similar to a palette. In work involving recovering the most plausible illuminant from pixel values in an image [9] it was found beneficial to utilize this kind of color signature. Here, the step of binarizing the histogram not only reduces the computational burden, since true chromaticity histograms need not be computed, but also has the effect of producing far fewer negatives in the compressed histogram. Finally, we found that one further step could substantially improve the energy compaction of the representation: we carry out a 16*16 Discrete Cosine Transform (DCT) on the compressed 16*16 histogram. After zigzag ordering, we keep 21 DCT coefficients.

Since every image now lives in approximately the same lighting, we can in fact precompute a basis for the DCT 21-vectors, offline, which can then be reused for any new image or video. Here we determine a basis set by the Singular Value Decomposition (SVD) of the DCT 21-vectors. We found that 12 components in the new basis represent that entire DCT vector very well and that energy compaction worked better using a spherical chromaticity, rather than the usual linear one.

Thus the method we set out here is to precompute a set of basis vectors, once and for all, and then form the 12-vector coefficients for any video frame with respect to this basis. Then keyframe extraction by clustering can be carried out very efficiently, using only 12-component vectors.

A keyframe is extracted from each of segmented scenes in a video. We use a hierarchical clustering scheme to segment a video into a sequence of scenes [5]. This method executes a bottom-up multi-stage merging process where only adjacent frames or frame groups are merged by calculating their *L2* distance, as we wished to maintain the temporal order. A threshold of distance is assigned to determine the final clusters, and each of those clusters corresponds to a scene. Finally, a keyframe is extracted from the medoid of each cluster.

# 3. VIDEO CLASSIFICATION BASED ON HMM

We use a hidden Markov model based method for video content classification based on visual and temporal features. The result is a HMM-based video classification method.

## 3.1 Hidden Markov Model

In an HMM, there are a finite number of states and the HMM is always in one of those states. At each clock time, it enters a new state based on a transition probability distribution depending on the previous state. After a transition is made, an output symbol is generated based on a probability distribution, depending on the current state.

In the formal definition of HMM, the *hidden states* are denoted $Q = \{q_1, q_2, \ldots, q_N\}$, where N is the number of states and the *observation symbols* are denoted $V = \{v_1, v_2, .., v_M\}$, where M is the number of observation symbols. The *state transition probability distribution* between states is represented by a matrix $A = \{a(i, j)\}$, where $a(i, j) = \Pr(q_j \text{ at } t+1 \mid q_i \text{ at } t)$, and the *observation symbol probability distribution* is represented by matrix $B = \{b_j(k)\}$, where $b_j(k)$ is the probability of generating observation $v_k$ when the current state is $q_j$. *Initial state distribution* denoted by $\pi = \Pr(q_i \text{ at } t = 1)$ contains the probabilities of the model being in every hidden state *i* at time *t*=1 that is the start point for a HMM.

A HMM is always represented by $\lambda = (A, B, \pi)$. We constructed four HMMs, corresponding to news, commercials, football game, and basketball game, respectively.

## 3.2 HMM Process

The HMM process consists of two phases, *viz.* training and classification. Figure.1 shows the training process for the basketball game HMM and classification process for a given video clip.

### *3.2.1 Training:*

The HMM training step is essentially to create a set of hidden states *Q* and a state transition probability matrix *A* for each video

topic category. The process of training a HMM for basketball videos is illustrated in Fig.1. The other three HMMs are trained in the same way.

We first summarize all videos in the basketball game training set to extract chromaticity signatures of keyframes. Then we cluster these signatures and take the medoids of resulting clusters as hidden states a video topic category. Here we use the CLARANS clustering algorithm [10]. This algorithm is an improved k-medoids clustering algorithm based on randomized search, which is effective and efficient in spatial data mining with large data sets.

The state transition probability matrix includes the probability of moving from one hidden state to another. There are at most $M^2$ (M is the number of states) transitions among the hidden states. Since each of clusters obtained from the above step corresponds to a hidden state and each keyframe in these clusters corresponds to a set of frames, we calculate the probabilities based on the number of frames falling into these clusters and the number of frames temporally transiting between clusters.

### *3.2.2 Classification:*

In the classification phase, given a target video, we make an observation sequence and feed it into HMMs as an input. By evaluating the probability for each HMM, the target video is assigned into a topic category with the highest probability of the HMM. Figure.1 shows the classification process for a given video clip.

For the observation sequence of the target video, we first summarize the target video and extract a set of keyframes in time order and take these keyframes as observation symbols. We then build a *temporal and keyframe-based summarized video sequence* (TSV) that is replicating each keyframe a number of times equaling the number of frames represented by the keyframe in the video sequence, and order these keyframes by time. In this way, a temporal feature can be maintained in the resulting sequence.

We also need to compute the observation symbol probability matrix *B* for each HMM, containing the output probabilities of the observation symbols given a particular hidden state. We compute these probabilities by the *inverse L2 distance* that is larger distance between an observation frame vector and a state vector, less probabilities of the observation frames belonging to the state. We have to calculate this matrix in the classification phase because the observation symbols of video keyframes are in an infinite set so that there is no way to train it in advance. The rationale behind the use of the distance for the probability is that it is visual distance or similarity that stands for the relationship between observations and states in this video case.

We use the *Forward* algorithm to calculate a probability for each HMM, and thus choose the video type with the most probable HMM. The *Forward* algorithm first defines a *partial probability*,

$\partial(t,j)$= Pr( observation symbol | hidden state is j ) × Pr(all paths to state j at time t),

which is the probability of reaching an intermediate state, then recursively calculates the probability of observing a sequence given a HMM,

$$\partial(t+1,\, j) = b(t+1,\, j) \cdot \sum_i (\partial(t, i) \cdot a(i,\, j)) \quad from\ \ t\ \ to\ \ t+1$$

and

$$\partial(1,\, j) = b(1,\, j) \cdot \pi(\, j)\ at\ \ t = 1\,.$$

Finally we calculate the probability for each HMM with the sum of partial probabilities of reaching every state at the end moment. The forward algorithm is, in effect, based upon the lattice structure shown in Fig. 2. The key is that there is only *N* states (nodes at each time slot in the lattice), all the possible state sequences will remerge into these *N* nodes, no matter how long the observation video sequence is.

## 4. EXPERIMENT RESULTS

We evaluate our classification method by classifying four types of TV programs: news reports, commercials, live basketball games, and live football games. We collected 100 video clips of 5 minutes each from TV broadcasting as the training set for each video category. Another set of 30 video clips for each category was used as the category's testing set. We assume that the input video clips always belong to one of the four categories of TV program.

Since there is no simple theoretical way to choose the number of states for each video category, we had to greedily try different numbers of states. From the method, we know the basketball game HMM classifier gives best performance when the state number is 6. In the same way, we get 7 states for news report, 12 for commercial states, and 5 for football games.

Table 2 gives the *state transition probability matrix* of the basketball game HMM with 6 states. We see that diagonal items are much higher than the rest because every state stays at its state for a period of time and then goes to some other state. In contrast, some items are 0 or very low probabilities, and we can think of these states as representing the same semantic scenes but having no temporal relationship with each other.

Table 1 gives the classification results using four HMMs for the four video categories. From this table we observe that the classifier can accurately identify basketball games and football games, but the separation of commercials from news reports is somewhat less successful, although still impressive. The reason is that these categories contain much chromatic information and the state duration is usually short in their models.

## 5. CONCLUSION AND FUTURE WORK

In this paper, we have described a video content classifier based on HMM using chromaticity signatures from video summarization and their temporal relationship. The video characterization and summarization method represents the video as a series of compressed chromaticity signatures. The HMM process use these signatures and takes advantage of the temporal feature to train HMMs and evaluate the probability of the given video being in one of the four categories of TV program.

Since video includes various visual features, we plan to explore the issue of HMM classification in terms of other attributes such as object information and investigate methods to extract a temporal feature for those attributes, for HMM processing.

## 6. REFERENCES

[1] A.G. Hauptmannn and M.J. Witbrock. Story segmentation and detection of commercials in broadcast news video. *Proceedings of Advances in*

*Digital Libraries Conference*, Santa Barbara, CA., April 22-24, 1998.

[2] Wensheng Zhou, Asha Vellaikal and C. C. Jay Kuo. Rule-based video classification system for basketball video indexing. *Proceedings on ACM multimedia 2000 workshops*, 2000, Pages 213 - 216

[3] G.Wei,L.Agnihotri, and N. Dimitrova. TV program classification based on face and text Processing. *IEEE multimedia and Expo 2000*, New York, July 2000.

[4] M.S.Drew, J.Wei, and Z.N.Li. Illumination-invariant color object recognition via compressed chromaticity histograms of color-channel-normalized images. In *ICCV98*, pages 533-540, IEEE, 1998.

[5] Mark S. Drew and James Au. Video keyframe production by efficient clustering of compressed chromaticity signatures. *ACM Multimedia '00*, pp.365-368, November 2000

[6] L.R.Rabiner and B.H.Juang. A tutorial on Hidden Markov Models. *IEEE ASSP Magazine*. pp4-15, January 1986.

[7] Nevenka Dimitrova, Lalitha Agnihotri and Gang Wei . Video classification based on HMM using text and faces. *European Conference on Signal Processing*, Finland, September 2000

[8] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. K. Wong. Integration of multimodal features for video classification based on HMM", *1999 IEEE Third Workshop on Multimedia Signal Processing*, pp. 53 - 58, Copenhagen, Denmark, Sept 13 - 15, 1999

[9] G.D.Finlayson, P.M.Hubel, and S.Hordley. Colorur by correlation. In Fifth Color Imaging Conf., page 6-11, 1997.

[10] R. Ng and J. Han, Efficient and effective clustering method for spatial data mining, *Proc. of 1994 Int'l Conf. on Very Large Data Bases (VLDB'94)*, Santiago, Chile, September 1994, pp. 144-155

**Table 1. Classification results (unit: 100%)**

| Result / Expectation | News | Commercial | Basketball | Football |
|---|---|---|---|---|
| **News** | 80.0 | 13.3 | 0 | 6.7 |
| **Commercial** | 23.3 | 66.7 | 6.7 | 3.3 |
| **Basketball** | 3.3 | 3.3 | 93.3 | 0 |
| **Football** | 0 | 3.3 | 0 | 96.7 |

**Table 2. State transition probability matrix of basketball game's HMM with 6 states**

| States *to* / *from* | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **1** | 0.9452 | 0.0142 | 0.0073 | 0.0002 | 0.0209 | 0.0122 |
| **2** | 0.0301 | 0.8712 | 0.0316 | 0.0210 | 0.0460 | 0.0001 |
| **3** | 0.0187 | 0.0258 | 0.8405 | 0.0637 | 0.0480 | 0.0033 |
| **4** | 0.0000 | 0.0450 | 0.0541 | 0.8166 | 0.0118 | 0.0725 |
| **5** | 0.0733 | 0.0475 | 0.0835 | 0.0216 | 0.7534 | 0.0207 |
| **6** | 0.1127 | 0.0004 | 0.0750 | 0.0947 | 0.0681 | 0.6491 |

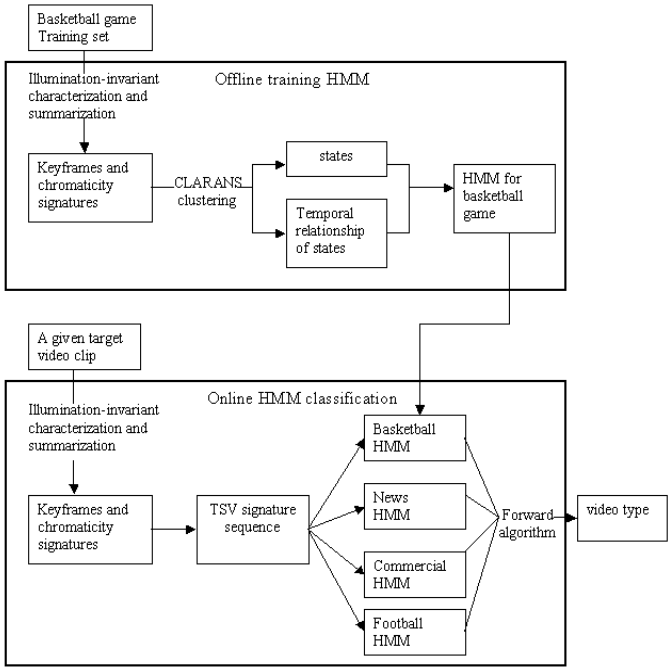


**Figure 1. HMM training and classification algorithm**

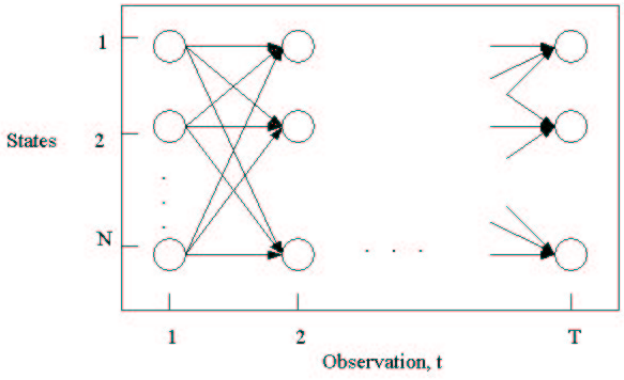


**Figure 2. Implementation of the computation of $\partial_t(i)$ in terms of a lattice of observations *t*, and states *i*.**