Hidden Markov Models for Video Skim Generation

Sergio Benini, Pierangelo Migliorati and Riccardo Leonardi
DEA-SCL, University of Brescia, Via Branze 38, 25123, Brescia, Italy

Abstract

In this paper we present a statistical framework based on hidden Markov models (HMMs) for video skimming. A chain of HMMs is used to model subsequent story units: HMM states represent different visual-concepts, transitions model the temporal dependencies in each story unit, and stochastic observations are given by single shots. The skim is generated as an observation sequence, where, in order to privilege more informative segments for entering the skim, dynamic shots are assigned higher probability of observation. The effectiveness of the method is demonstrated on a video set from different kinds of programmes, and results are evaluated in terms of metrics that assess the content representational value of the obtained video skims.

1. Introduction

In the last years, with the proliferation of digital TV broadcasting, dedicated internet websites, private recording of home video, a large amount of video information has been made available to end-users. In this scenario, video abstraction techniques may represent a key component of a practical video-content management system. By watching at a condensed video, a viewer may be able to assess the relevance of a programme before committing time, thus facilitating typical tasks such as browsing, organizing and searching content. Regarding videos which are produced according to a script, such as movies, news and cartoons, two types of video abstraction have been investigated so far, namely *video static summarization* and *video skimming*. The first one is a process that selects a set of salient key-frames to represent content in a compact form and present it to the user as a static programme preview. Video skimming instead, also known as *video dynamic summarization*, tries to condense the original video in the more appealing form of a shorter video clip. The generation of a skim can be viewed as the process of selecting and gluing together proper video segments under some user-defined *constraints* and according to given *criterions*. End-user *constraints* are usually defined by the time committed by the user to watch the skim, which in the end determines the final skim ratio. On the other hand, skimming *criterions* used to select video segments range from the exploitation of the hierarchical organization of video in scenes and shots as in [5], the use of motion information [4], or the insertion of audio, visual and text markers [7].

In this paper, in order to derive the skim, we propose to combine the information deriving from the story structure with the characterization of the motion dynamics of shots. Through partial decoding of the compressed stream, we compute a motion descriptor which inherently estimates the contribution of each shot in term of "content informativeness" and determines whether the shot will be included in the final skim. The shot sequence which forms the skim is obtained as a series of observations of a *HMM* chain, where each *HMM* models the structure of a semantic scene. In the past *HMM* has been successfully applied to different domains such as speech recognition, genome sequence analysis, etc. For video analysis, *HMMs* have been used to distinguish different genres [8], and to delineate high-level structures of soccer games [9]. In this work instead, *HMMs* are used as a unified statistical framework to represent visual-concepts and to model the temporal dependencies in story units with the aim of video skimming.

The paper is organized as follows. Section 2 presents the criterions adopted to realize the skim. In Section 3 we characterize the general dynamics of shots by a motion descriptor. Section 4 describes how to model each story unit by a *HMM*. In Sections 5 and 6 the video skims are generated and evaluated, while in Section 7 conclusions are drawn.

2 Skimming criterions

We propose that the time allocation policy for realising a skim should take into account the following criterions:
"Coverage": the skim should include all the parts of the story structure into the synopsis (*i.e.,* all the story units);
"Representativeness": each story unit should be represented proportionally to its duration in the original video;
"Structure informativeness": the information which is introduced by the film editing process, especially that conveyed by the shot patterns inside story units (*e.g.,* dialogues,

progressive scenes, *etc.*) should be included in the skim; *"Content informativeness"*: to represent each story unit, the most "informative" video segments should be preferred. In the next, we start investigating the *content informativeness* of shots, by relying on a measure of the motion activity.

3 Motion activity analysis

As stated in [3] the intensity of motion activity in a video segment is in fact a measure of "how much" the content is changing. Thus motion activity can be interpreted as a measure of the "entropy" (in a wide sense) of a video segment. We characterize the motion activity of shots by extracting the motion vector *(MV)* field of *P*-frames from the compressed *MPEG* stream, with low computational cost.

The extracted raw *MV* field turns out to be normally rough and erratic; however, after being properly filtered, the *MVs* can be useful to characterize the general motion dynamics of a sequence. The filtering process applied includes first removing the *MVs* next to image borders which tend to be unreliable, then using a texture filter, followed by a median filter. The texture filter is needed since, in the case of low-textured uniform areas, the correlation methods used to estimate motion often produce spurious *MVs*.

3.1 Motion Intensity

Of course, the perceived motion activity in a video is higher when the objects in the scene move faster. In this case the magnitudes of the *MVs* of the macro-blocks (*MBs*) that make up the objects are significant, and one simple measure of motion intensity can be extracted from the *P*-frame by computing the mean μ_P of the magnitudes of motion vectors belonging to inter-coded *MBs* only (intra-coded *MBs* have no *MVs*). However, most of the perceived intensity in a video is due to objects which do not move according to the uniform motion of camera. Thus, a good *P*-frame-based measure of motion intensity is given by the standard deviation σ_P of the magnitudes of motion vectors belonging to inter-coded *MBs*.

This measure can be also extended to characterize the motion intensity of a shot S, by averaging the measures obtained on all the *P*-frames belonging to the shot. *MPEG7 Motion Activity* descriptor [3] is also based on a quantized version of the standard deviation of *MVs* magnitudes. For our purposes, to each shot S is assigned its motion intensity value $\mathcal{MI}(S)$ in its not-quantized version. This value $\mathcal{MI}(S)$ tries to capture the human perception of the "intensity of action" or the "pace" of a shot, by considering the overall intensity of motion activity in the shot itself (without distinguishing between the camera motion and the motion of objects present in the scene). Since this is in fact a measure of "how much" the content of a video is changing, it can be interpreted as a measure of the "entropy" of the video segment, and can be used for summarization purposes.

4 HMM for LSU representation

In [10] it is shown that after the removal of cut-edges of a *Scene Transition Graph (STG)*, each connected sub-graph well represents a *Logical Story Unit (LSU)*, *i.e.,* "a sequence of contiguous and interconnected shots sharing a common semantic thread", which is the best computable approximation to a semantic scene [2]. In particular sub-graph nodes are clusters of visually similar and temporally close shots, while edges between nodes give the temporal flow inside the *LSU*, as shown in Figure 1.

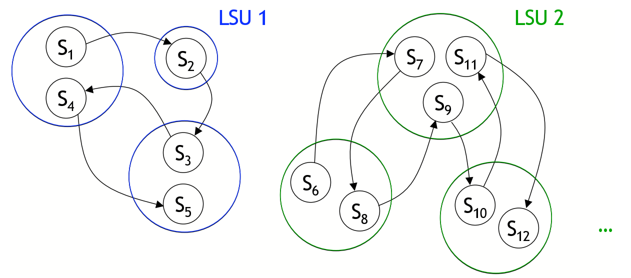

Figure 1. Logical Story Units in a STG.

Starting from the *STG* representation, each *LSU* can be equivalently modeled by a *HMM*. This is a discrete state-space stochastic model which works well for temporally correlated data streams, where the observations are a probabilistic function of a hidden state [6]. Such a modeling choice is supported by the following considerations ([9]):
i. Video structure can be described as a discrete state-space, where each state is a conveyed *concept* (*e.g., "man face"*) and each state-transition is given by a change of concept;
ii. The *observations* of concepts are stochastic since video segments seldom have identical raw features even if they represent the same concept (*e.g.,* more shots showing the same *"man face"* from slightly different angles);
iii. The sequence of concepts is highly correlated in time, especially for scripted-content videos (movies, etc.) due to the presence of editing effects and typical shot patterns inside scenes (*i.e.,* dialogues, progressive scenes, etc.).

For our aims *HMM* states representing concepts will correspond to distinct clusters of visually similar shots; state transition probability distribution will capture the shot pattern structure of the *LSU*, and shots will constitute the observation set (as shown in Figure 2).

4.1 HMM definition

Formally, a *HMM* representing an *LSU* is specified by:
• N, *the number of states.* Although the states are hidden,

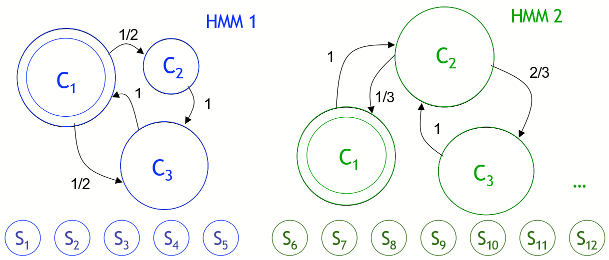

Figure 2. LSU modeling by HMMs.

in practical applications there is often some physical significance associated to the states. In this case we define that each state corresponds to a distinct node of a *STG* subgraph: each state is one of the N clusters of the *LSU* containing a number of visually similar and temporally close shots. We denote states as $C = \{C_1, C_2, \dots, C_N\}$, and the state at time t as q_t.

• *M, the number of distinct observation symbols.* The observation symbols correspond to the output of the system being modeled. In this case, each observation symbol $S = \{S_1, S_2, \dots, S_M\}$ is one of the M shots of the video.

• $\Delta = \{\delta_{ij}\}$, *the state transition probability distribution*:

$$\delta_{ij} = P[q_{t+1} = C_j | q_t = C_i], \qquad 1 \le i, j \le N$$

Transition probabilities are computed as the relative frequency of transitions between clusters in the *STG*, *i.e.,* δ_{ij} is given by the ratio of the number of edges going from cluster C_i to C_j to the total number of edges departing from C_i.

• $\Sigma = \{\sigma_j(k)\}$, *the observation symbol distribution*, where

$$\sigma_j(k) = P[S_k \text{ at } t | q_t = C_j], \quad 1 \le j \le N, 1 \le k \le M$$

We define the observation symbol probability in state C_j, that is $\sigma_j(k)$, as the ratio of the motion intensity of the shot S_k to the *total motion intensity* of the cluster, that is:

$$\sigma_j(k) = \begin{cases} \frac{\mathcal{MI}(S_k)}{\mathcal{MI}(C_j)} & \text{if } S_k \in C_j \\ 0 & \text{otherwise ,} \end{cases}$$

where $\mathcal{MI}(C_j)$ is defined as the sum of all the motion intensity of the shots belonging to cluster C_j.

• $\pi = \{\pi_i\}$, *the initial state distribution*, where:

$$\pi_i = P[q_1 = C_i], \qquad 1 \le i \le N .$$

In order to preserve the information about the entry point of each *LSU*, $\pi_i = 1$ if the cluster C_i contains the first shot of the *LSU*, otherwise $\pi_i = 0$. Therefore a complete specification of an *HMM* requires two model parameters (N and M), the observation symbols S, and the probability distributions Δ, Σ and π. Since the set $S = \{S_1, S_2, \dots, S_M\}$ is common to all the *HMMs*, for convenience, we can use the compact notation $\Lambda = (\Delta, \Sigma, \pi, N)$ to indicate the complete parameter set of the *HMM* representing an *LSU*.

5 Stochastic skim generation

In order to generate an informative skim, the following solutions have been adopted to fulfill all the skimming criterions stated in Section 2.

Coverage: Since the skim should include all the semantically important story units, each detected *LSU* participates to it (where the skim ratio is subject to a minimal value).

Representativeness: Let $l_1, l_2, \dots, l_n$ be the lengths of the n *LSUs* that compose the original video. Then in the skim, for each λ_i, a time slot of length ξ_i is reserved, where ξ_i is proportional to the duration of λ_i in the original video.

Structure informativeness: In order to include in the synopsis the information conveyed by the shot patterns inside the story units, a skimmed version of each *LSU* λ can be generated as an observation sequence of the associated *HMM*, Λ, that is:

$$O = O_1 O_2 \cdots ,$$

where each observation O, is one of the symbols from S. The sequence is generated as follows:

1. Choose the initial state $q_1 = C_i$ according to the initial state distribution π. Set $t = 1$;
2. While (total length of already concatenated shots) $<$ (time slot ξ assigned to the current *LSU*)
   (a) Choose $O_t = S_k$ according to the symbol probability distribution in state C_i, *i.e.*, $\sigma_i(k)$;
   (b) Transit to a new state $q_{t+1} = C_j$, according to the state transition probability for state C_i, *i.e.,* δ_{ij};
   (c) Set $t = t + 1$;

The above procedure is then repeated for all *LSUs*. Finally, all the obtained sequences of observed shots are concatenated in order to generate the resulting skim.

Content informativeness: In order to privilege the more “informative” shots, the observation symbol probability distribution Σ depends on the shot motion intensity. In particular the higher is the motion present in a shot S_k of the cluster C_j, the higher will be $\sigma_j(k)$, *i.e.,* S_k will be more likely chosen for the skim. Since motion activity can be interpreted as a measure of the “entropy” of a video segment, by assigning higher probability of observation to more dynamic shots, we privilege “informative” segments for the skim generation. At the same time, we avoid to discard *a-priori* low-motion shots, that can be chosen as well for entering the skim, even if with lower probability. Moreover, once that one shot is chosen for the video skim, it is removed from the list of candidates for further time slots, at least until all shots from the same cluster are employed too. This prevents the same shot from repetitively appearing in the same synopsis, and at the same time it favorites the presence of low-motion shots, if the desired skim ratio is big enough. Therefore, as it should be natural, in very short skims, "informative" shots are likely to appear first, while for longer skims, even less "informative" shots can enter the skim later on.

6 Performance evaluation

To quantitatively investigate the performance of video skimming, we carried out some experiments using the video sequences in Table 1 for a total time of about four hours of video and more than two thousands shots.

Table 1. Video data set.

No.	Video (genre)	Length	Shots
1	*Portuguese News (news)*	47:21	476
2	*Notting Hill (movie)*	30:00	429
3	*A Beautiful Mind (movie)*	17:42	202
4	*Pulp Fiction (movie)*	20:30	176
5	*Camilo & Filho (soap)*	38:12	140
6	*Riscos (soap)*	27:37	423
7	*Misc. (basket/soap/quiz)*	38:30	195
8	*Don Quixotte (cartoon)*	15:26	188

For the evaluation of the skims, the method and the criterions of *"informativeness"* and *"enjoyability"* adopted in [5] have been used. *Informativeness* assesses the capability of the statistical model of maintaining content, coverage, representativeness and structure, while reducing redundancy. *Enjoyability* instead assesses the performance of the motion analysis in selecting perceptually enjoyable video segments for the skim. Starting from the *LSU* segmentation results we presented in [1], we generated eighteen dynamic summaries with their related soundtracks: for each video two associated skims have been produced, one with 10% of the original video length and the other with the 25%. Ten students assessed the quality of the skims by watching first the 10% one, then the 25%, and finally the original video. After watching a skim, each student assigned two scores ranging from 0 to 100, in terms of *informativeness* and *enjoyability*. Then students were also requested to give scores to the original videos in case they thought that these were not 100% informative or enjoyable. On this basis, after watching the original, the students were also given the chance to modify the scores assigned before to the associated skims. Finally the scores assigned to skims have been normalized to the scores given to the original video.

In these experiments, average normalized scores for *enjoyability* are around 72% and 80%, respectively, for video skims of 10% and 25% length. Regarding *informativeness*, average normalized scores are around 68% and 81%, respectively. These results are comparable with results presented in most recent works on video skims [5], but they have been obtained on a larger set of videos of different genres. Moreover, since the skim generation does not take into account the original shot order (*i.e.,* in the skim a shot which is later in the original video can appear before another shot which is actually prior to it, as it sometimes happens in commercial trailers), nevertheless the obtained results suggest that the skim preserves its informativeness and that the viewer is not particularly disturbed if some shots are shown in non sequential order.

7 Conclusions

In this paper a method for video skim generation has been proposed. This technique is based on a previous high-level video segmentation and on the use of *HMMs*. The final skim is a sequence of shots which are obtained as observations of the *HMMs* corresponding to the story units, and by a motion measure which roughly estimates each shot "informativeness". The effectiveness of the proposed solution has been demonstrated in terms of informativeness and enjoyability on a large video set coming from different genres.

References

[1] S. Benini, A. Bianchetti, R. Leonardi, and P. Migliorati. Video shot clustering and summarization through dendrograms. In *Proc. of WIAMIS'06*. Incheon, South Korea, 19-21 April 2006.

[2] A. Hanjalic, R. L. Lagendijk, and J. Biemond. Automated high-level movie segmentation for advanced video retrieval systems. *IEEE Trans. on CSVT*, 9(4), Jun 1999.

[3] S. Jeannin and A. Divarakan. MPEG7 visual motion descriptors. *IEEE Trans. on CSVT*, 11(6), Jun 2001.

[4] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *Proc.* 10^{th} *ACM Conf. on MM.*, pages 533–542. Juan Les Pins, France, Dec 2002.

[5] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang. Video summarization and scene detection by graph modeling. *IEEE Trans. on CSVT*, 15(2):296–305, Feb 2005.

[6] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[7] M. A. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding. In *Proc. of IEEE Int. Work. on Content-Based Access Image Video Data Base*, pages 61–67, Jan 1998.

[8] Y. Wang, Z. Liu, and J.-C. Huang. Multimedia content analysis using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(11):12–36, Nov 2000.

[9] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with hidden markov model. In *Proc. of ICASSP'02*. Orlando, Florida, USA, May 2002.

[10] M. M. Yeung and B.-L. Yeo. Time-constrained clustering for segmentation of video into story units. In *Proc. of ICPR'96*, volume III-vol. 7276, page 375. Vienna, Austria, Aug 1996.