A STATISTICAL FRAMEWORK FOR VIDEO SKIMMING BASED ON LOGICAL STORY UNITS AND MOTION ACTIVITY

Sergio Benini, Pierangelo Migliorati, Riccardo Leonardi

DEA-SCL, University of Brescia, Via Branze 38, 25123, Brescia, Italy
Tel: +39 030 3715528, Fax: +39 030 380014, E-mail:{firstname.lastname}@ing.unibs.it

ABSTRACT

In this work we present a method for video skimming based on hidden Markov Models (*HMMs*) and motion activity. Specifically, a set of *HMMs* is used to model subsequent logical story units, where the *HMM* states represent different visual-concepts, the transitions model the temporal dependencies in each story unit, and stochastic observations are given by single shots. The video 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 measure the content representational value of the obtained video skims.

1. INTRODUCTION

The proliferation of dedicated internet websites, digital TV broadcasting, private recording of home video, has provided the end-users with a large amount of video information. Nevertheless, this massive proliferation in the *availability* of digital video has not been accompanied by a parallel increase in its *accessibility*. In this scenario, video summarization 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 video-content.

For unscripted-content videos such as sports and home-videos, where the events happen spontaneously and not according to a given script, previous work on video abstraction mainly focused on the extraction of *highlights*. Regarding scripted-content videos, that are 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 *static video summarization* and *video skimming*.

The first one is the process of selection of a reduced set of salient key-frames to represent the content in a compact form and to present it to the final user as a static programme preview. On the other hand, video skimming, also known as *dynamic video summarization*, tries to condense the original video in the more appealing form of a shorter video clip.

The generation of a video 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*. The end-user *constraints* are usually defined by the time committed by the user to watch the skim, which in the end determines the final skimming 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 [1], the use of motion information [2], or the insertion of audio, visual and text markers [3].

In this paper the video skim is generated by combining the information deriving from the story structure with the characterization of the motion activity of the video shots. More specifically, we compute a motion descriptor which inherently estimates the contribution of each shot in term of "content informativeness" and determines whether or not the shot would be included into the final skim. The shot sequence which forms the skim is then obtained as a series of observations of a *HMM* chain, where each *HMM* try to model the structure of each semantic scene, so capturing the "structure informativeness" of the video.

In the past *HMMs* have been successfully applied to different domains such as speech recognition, handwriting recognition, or genome sequence analysis. For video analysis, *HMMs* have been used to distinguish different genres [4], and to delineate high-level structures of soccer [5] and tennis games [6]. In this work instead, *HMMs* are used as a unified statistical framework to represent visual-concepts and to model the temporal dependencies in the video story units, with the aim of effective video skimming.

The rest of the paper is organized as follows. Section 2 presents the general criterions adopted to realize the skim, whereas in Section 3 we estimate the intrinsic dynamics of the video shots by a suitable motion activity descriptor. Section 4 describes how each logical story unit is modeled by a *HMM*. In Sections 5 and 6 the video skims are generated and evaluated, respectively. Concluding remarks are given in Section 7.

2. GENERAL CRITERIONS

Since a skimming application should automatically shorten the original video while preserving the important and informative content, we propose that the time allocation policy for realising a video skim should take into account the following criterions:

- *"Coverage"*: the skim should include all the elements of the story structure into the final synopsis (*i.e.,* all the story units);
- *"Representativeness"*: each story unit should be represented in the skim 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 into the skim;
- *"Content informativeness"*: to represent each story unit, the most "informative" video segments should be preferred.

In the next paragraph, we start investigating the *content informativeness* of video shots, by relying on a measure of the motion activity.

3. MOTION ACTIVITY DESCRIPTOR

The intensity of motion activity is a subjective measure of the perceived intensity of motion in a video segment. For instance, while an *"anchorman"* shot in a news program is perceived by most people as a "low intensity" action, a *"car chasing"* sequence would be viewed by most viewers as a "high intensity" sequence.

As stated in [7] the intensity of motion activity in a video segment is in fact a measure of "how much" the content of a video is changing. Motion activity can be therefore interpreted as a measure of the "entropy" (in a wide sense) of a video segment. We characterize the motion activity of video shots by extracting the motion vector *(MV)* field of *P*-frames (see Figure 1) directly from the compressed *MPEG* stream, thus allowing low computational cost. By characterizing its general motion dynamics, we characterize the amount of visual information conveyed by the shot.

3.1. Filtering of the Motion Field

The raw *MV* field extracted turns out to be normally rough and erratic, and not suitable for tasks such as accurately segmenting moving objects. However, after being properly filtered the *MVs* can be very useful to characterize the general motion dynamics of a sequence. The filtering process applied

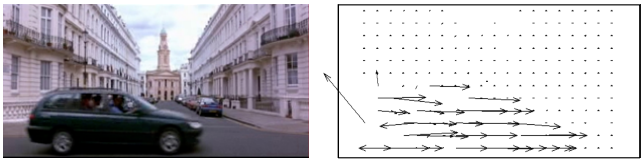

Fig. 1. A decoded *P*-frame and its motion vector field.

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*. After having filtered the motion vectors on texture criterion, a median filtering is used to straighten up single spurious vectors such as those that could still be present close to borders.

3.2. Evaluation of the Motion Intensity

In general, 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 the video 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*.

The measure σ_P, can be also extended to characterize the motion intensity $\mathcal{MI}(S)$ of a shot S, by averaging the measures obtained on all the *P*-frames belonging to that shot. *MPEG7 Motion Activity* descriptor [7] is also based on a quantized version of the standard deviation of *MVs* magnitudes. For our purposes, 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 video segment, by considering the overall intensity of motion activity in the shot itself (without distinguishing between the camera motion and the motion of the objects present in the scene). Since this is in fact a measure of "how much" the content of a video segment is changing, it can be interpreted as a measure of the "entropy" of the video segment, and can be used also for summarization purposes.

4. LSU REPRESENTATION

In [8] Yeung *et al.* shown that in a *Scene Transition Graph (STG)*, after the removal of cut-edges, each connected subgraph well represents a *Logical Story Unit (LSU)*, *i.e.,* "a se-

quence of contiguous and interconnected shots sharing a common semantic thread", which is the best computable approximation to semantic scene [9]. 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 2.

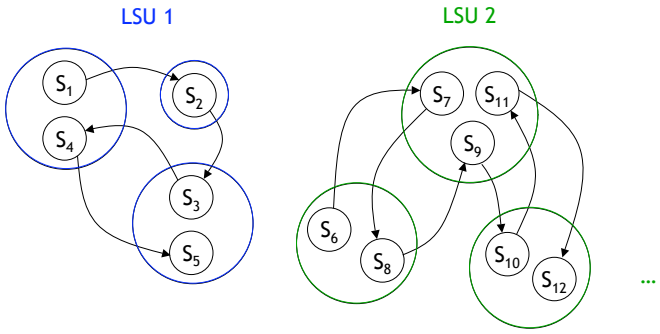

Fig. 2. After the removal of *cut-edges*, each connected sub-graph of the *STG* represents a Logical Story Unit.

Starting from the *STG* representation, each *LSU* can be equivalently modeled by a *HMM*. This is a discrete state-space stochastic model which works quite well for temporally correlated data streams, and where the observations are a probabilistic function of a hidden state [10]. Such a modeling choice is supported by the following considerations (see [5]):
i) Video structure can be described as a discrete state-space, where each state is a conveyed *concept* (*e.g., "man talking"*) 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 talking"* 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, as described in the next Section, *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 3).

4.1. *HMM* definition

Formally, a *HMM* representing an *LSU* is specified by:
• N, *the number of states*. Although the states are hidden, 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* sub-graph: each state is one of the N clusters of the *LSU* containing a number of visually similar and temporally close shots. We denote

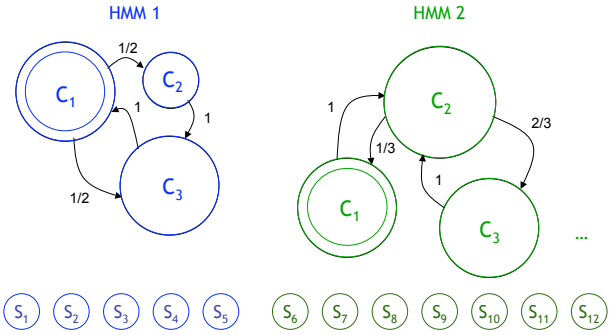

Fig. 3. *LSUs* of Figure 2 are equivalently modeled by *HMMs*.

states as $C = \{C_1, C_2, \ldots, C_N\}$, and the state at time t as q_t.
• M, *the number of distinct observation symbols*. The observation symbols correspond to the physical output of the system being modeled. In this case, each observation symbol $S = \{S_1, S_2, \ldots, S_M\}$ is one of the M shots of the video.
• $\Delta = \{\delta_{ij}\}$, *the state transition probability distribution*:

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

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

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

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, that is:

$$\mathcal{MI}(C_j) = \sum_{S_h \in C_j} \mathcal{MI}(S_h) .$$

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

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

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 $\pi_i = 0$.

From the above discussion it can be seen that 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 $S = \{S_1, S_2, \ldots, S_M\}$ 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. SKIM GENERATION

In order to generate an informative skim which fulfills the criterions stated above, the following solutions have been adopted.

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

- *Representativeness:* Let $l_1, l_2, \ldots, l_n$ be the lengths of the n *LSUs* the original video has been segmented in. Then in the skim, for each *LSU* 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;
2. Set $t = 1$;
3. 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 *HMM* generated sequences of observed shots are concatenated in order to generate the resulting video 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 investigate the performance of the proposed video skimming method, we carried out some experiments using the video sequences described 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
9	*Music Show (music)*	10:00	122

For the evaluation the method the two criterions of *"informativeness"* and *"enjoyability"* adopted in [1] 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 skims.

Starting from the *LSU* segmentation results we presented in [11], 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% of the original length.

Ten students assessed the quality of these video skims from high to low skim ratio, *i.e.*, by watching first the 10% video, then the 25%, and finally the original video (100%). After watching a video, 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 videos were not 100% informative or enjoyable. On this basis, after watching

the original video, the students were also given the chance to modify the original scores assigned before to the two 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% skimming ratio. Regarding *informativeness* instead, average normalized scores are around 68% and 81%, respectively. These preliminary results are comparable with results presented in most recent works on video skims [1], but they have been obtained on a larger data set of video coming from 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, at least as long as the visual-content remains coherent.

7. CONCLUSIONS

In this paper we have proposed a method for the automatic generation of video skims based on motion activity and hidden Markov Models. The final skim is a sequence of shots which are obtained as observations of the *HMMs* corresponding to each story units. In particular the observation probability distribution for shots is determined by a motion activity measure which roughly estimates each shot contribution in terms of "content informativeness". The effectiveness of the proposed method has been demonstrated in terms of informativeness and enjoyability on a large video set coming from different genres.

8. REFERENCES

[1] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Trans. on CSVT*, vol. 15, no. 2, pp. 296–305, Feb 2005.

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

[3] 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*, Jan 1998, pp. 61–67.

[4] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Processing Magazine*, vol. 17, no. 11, pp. 12–36, Nov 2000.

[5] 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.

[6] E. Kijak, L. Oisel, and P. Gros, "Hierarchical structure analysis of sport videos using hmms," in *Proc. of ICIP'03*. Barcelona, Spain, September 2003, pp. 1025–1028.

[7] S. Jeannin and A. Divarakan, "MPEG7 visual motion descriptors," *IEEE Trans. on CSVT*, vol. 11, no. 6, Jun 2001.

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

[9] A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automated high-level movie segmentation for advanced video retrieval systems," *IEEE Trans. on CSVT*, vol. 9, no. 4, Jun 1999.

[10] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb 1989.

[11] 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.