RETRIEVAL OF VIDEO STORY UNITS BY MARKOV ENTROPY RATE

Sergio Benini, Pierangelo Migliorati, and Riccardo Leonardi

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

ABSTRACT

In this paper we propose a method to retrieve video stories from a database. Given a sample story unit, *i.e.,* a series of contiguous and semantically related shots, the most similar clips are retrieved and ranked. Similarity is evaluated on the story structures, and it depends on the number of expressed visual concepts and the pattern in which they appear inside the story. Hidden Markov Models are used to represent story units, and Markov entropy rate is adopted as a compact index for evaluating structure similarity. The effectiveness of the proposed approach is demonstrated on a large video set from different kinds of programmes, and results are evaluated by a developed prototype system for story unit retrieval.

Index Terms— Video retrieval, Logical Story Units *(LSU)*, Hidden Markov Model *(HMM)*, Markov entropy rate.

1. INTRODUCTION

The increasing growth of online video data, TV broadcasting, and private collections of personal videos, are contributing to make video retrieval an active research area. In spite of this proliferation of digital video, such increasing availability of this kind of information has not been accompanied by a parallel increase in its accessibility.

The possibility of performing specific tasks on video such as, for example, retrieving a particular clip from a movie, or browsing all the scenes with a favourite actor from a large digital library, is still a hidden dream for many multimedia consumers. These difficulties are in part due to the nature of video, since this kind of data probably remains unsuitable for most traditional forms of indexing, search and retrieval, which are usually either text-based or employing the query-by-example paradigm.

In order to face the issue of an efficient retrieval of the desired piece of video information, recent research work attempted to automatically index different scenes inside movies [1], to extract video storyboards, to summarize news broadcasts [2], to identify highlights in sports [3], and even to isolate single actions and emotions [4] inside videos.

Most traditional forms of indexing, search and retrieval are based on shots, *i.e.*, the basic video segments filmed in one single camera take. However, if we consider that there are usually several hundreds of shots for an hour long video, shot decomposition often leads to a segmentation that is too fine.

Thus significant research efforts are now directed towards detection and retrieval of the larger pieces of information, called *story units*, that are groups of contiguous shots considered the "best computable approximation of a semantic scene" [5]. Story units in video documents are extensively researched with respect to feature films and sitcoms (see [6]). However, detection of story boundaries alone is not enough. For retrieval, we are especially interested in some accompanying index which facilitate the access to video-content and can help in retrieving semantically similar units.

In this paper, we first define three types of information that can be helpful to index a story unit, that are the information related to the story *content, time and structure*, respectively. Then we propose a novel method to index and retrieve story units based only on their structure similarity, that is by comparing the number of expressed visual concepts and the pattern in which they appear inside the story. Fast indexing and retrieval are performed by a measure of the Markov entropy computed on the *HMM* chain modeling the video stories.

In the past *HMM* has been successfully applied to video analysis to distinguish different genres [7], to delineate high-level structures of soccer games [8] and to detect dialogue scenes [9]. In this work, as a further development of the model introduced in [10] to generate automatic video skims, *HMMs* are used to represent visual-concepts and their temporal dependencies, with the aim of fast retrieval of the story units showing similar structure.

The paper is organized as follows. Section 2 describe the nature of video stories and three types of information which a story conveys. Then the preliminary segmentation process into story units is briefly described in Section 3, whereas Section 4 presents how each story unit can be equivalently modeled by a *HMM*. In Section 5 the compact index based on Markov entropy for the description of story structure is given. Finally in Sections 6 and 7 retrieval results are evaluated, and concluding remarks are given, respectively.

2. INFORMATIVENESS IN STORY UNITS

A story unit, also called *Logical Story Unit (LSU)*, is "a sequence of contiguous and interconnected shots sharing a common semantic thread" [5], as shown in Figure 1.

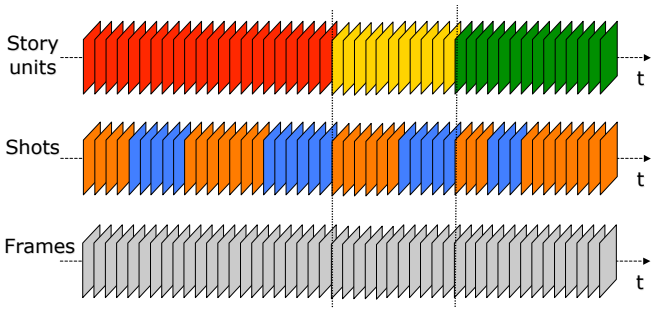

Fig. 1. Hierarchical decomposition of a video into shots and story units.

For a *LSU* the following properties apply [8]:

1. The story structure can be described as a discrete state-space $C = \{C_1, C_2, \ldots, C_N\}$, where each state is related to a conveyed *concept* (*e.g., "man talking"*) and each state-transition is given by a change of concept;
2. The *observations* $S = S_1 S_2 S_3 \ldots$ of concepts in a *LSU* are stochastic since video shots 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);
3. 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 story units. For example a sequence of concepts such as $C_1C_2C_1C_2$ probably represents a dialogue between two characters, while the sequence of concepts $C_1C_2C_3C_4$ may stand for the pattern of a progressive scene.

From these observations, we understand that a story unit conveys information on at least three different levels:

- **Content informativeness**. This is the information associated to the semantic concept represented by each shot (*e.g., "man talking"*). Many research attempts have been made so far to describe concepts using low-level audio-visual information on the available multi-modal channels [11];
- **Pace informativeness**. This is the information related to time: the rhythm given by the transitions between observations in the story unit (*i.e.,* the shot cut rate), for example, can distinguish a long rush of home-video from a car chasing scene in a feature movie;
- **Structure informativeness**. This is the information regarding the cardinality N of the concept set expressed in the story and the sequence pattern S in which they are observed (distinguishing for example a dialogue from a progressive scene).

In this work we want to retrieve similar story units from different videos relying only on structure informativeness. Although audio-visual content and time are usually considered more important than the structure, we aim to demonstrate the importance of this third type of information when retrieving stories.

3. STORY UNIT SEGMENTATION

The proposed method is based on the first automatic segmentation into story units, which adopts the procedure described in [12].

Based on the original proposal by Yeung *et al.* [1], the *Scene Transition Graph (STG)* is built as shown in Figure 2. Graph nodes are given by clusters of visually similar and temporally close shots, which are clustered on the basis of a vector quantization process on the *LUV* color space [12]. Edges between nodes represent the temporal flow between the shots of the *LSU*. A special type of edge between two nodes is called "cut-edge" [1], which, if removed, leads to the decomposition of the graph into two disconnected sub-graphs. After the removal of the cut-edges, each connected sub-graph well represents one *Logical Story Unit*.

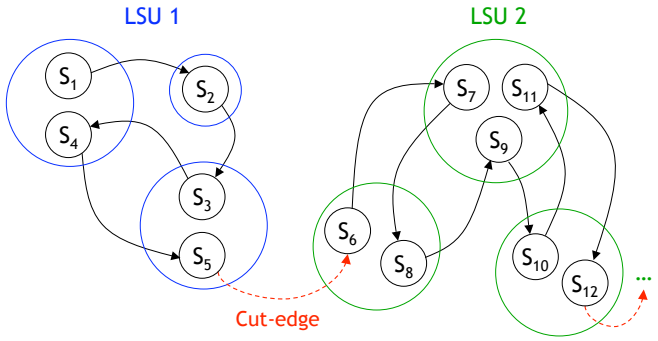

Fig. 2. After the removal of *cut-edges*, each connected sub-graph of the *STG* represents a Logical Story Unit.

4. HMM FOR LSU REPRESENTATION

Starting from the *STG* representation, each obtained *LSU* is modeled with an equivalent *HMM*, which is a discrete state-space stochastic model where observations are a probabilistic function of a hidden state, and which works well for temporally correlated data streams [13].

As shown in Figure 3, the *HMM states* represent distinct clusters of visually similar shots and stand for concepts; *state*

transition probability distribution captures the shot pattern in the *LSU*, and shots constitute the *observation set*. This model was formerly presented in [10] with the aim of automatically generating a video skim. In this work we adopt this model since on its basis, Markov entropy can be computed as a compact index for describing the story structure.

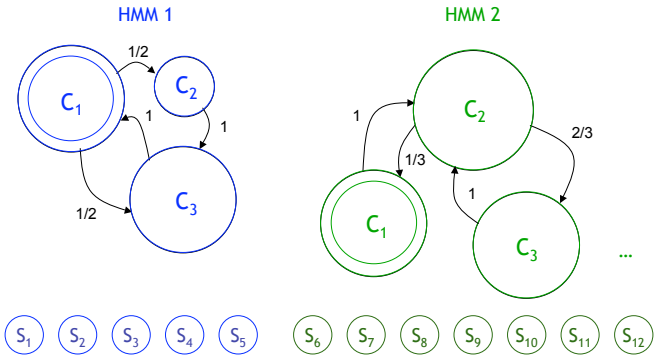

Fig. 3. *LSUs* of Figure 2 are equivalently modeled by *HMMs*.

More formally, a *HMM* representing a *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 concept, which is given by a distinct node of the *STG* sub-graph. Thus 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 $C = \{C_1, C_2, \ldots, C_N\}$, 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*:

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

  The observation symbol probability $\sigma_j(k)$ is null in case the shot S_k doesn't belong to C_j; on the contrary, if $S_k \in C_j$, then $\sigma_j(k)$ may represent the probability for a shot of being chosen (among all shots representing the same given concept) by the director during the editing process of video-making. An alternative *ad-hoc* definition of the observation symbol probability Σ can be found in [10] to preserve the content informativeness of a video skim.

- $\pi = \{\pi_i\}$, *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 arises that a complete definition of the *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 one *LSU*.

5. MARKOV ENTROPY RATE

Since the structure informativeness of a story unit depends on the number of expressed visual concepts (*i.e.,* N) and the pattern in which they appear (*i.e.,* Δ), for this method we can simplify the previous model and treat each story unit as a simple Markov chain $X = (\Delta, \pi, N)$, thus disregarding the value of Σ (since states are completely observable).

For any finite state Markov chain X with a unique stationary distribution, the stationary probability $\mu = \{\mu_i\}$ can be found by solving the equation

$$\mu\Delta = \mu \tag{1}$$

Then the Markov entropy rate for X is defined as in [14]:

$$H(X) = -\sum_{ij} \mu_i \cdot \delta_{ij} \cdot log(\delta_{ij}) \tag{2}$$

In general, the entropy rate of a dynamic process measures the uncertainty that remains in the next information produced by the process, given complete knowledge of the past. Thus it can be considered a natural measure of the difficulty to predict the process evolution.

When applied to a video story, entropy rate estimates the difficulty to predict the concept that will appear next. During a dialogue scene for example, it is quite easy to predict that, after character C_1, character C_2 will be shot, thus determining a low level of uncertainty. On the contrary in highly dynamic story units, such as progressive scenes, entropy rates will be higher, since it will be more difficult to predict the next concept. In fact, when progressive scene are forced to be clustered into a restricted number of concepts, the story

structure becomes more complex and the transition probabilities tend to be identically distributed. Examples of Markov chains with low and high entropy rates are given in Figure 4.a and Figure 4.b, respectively.

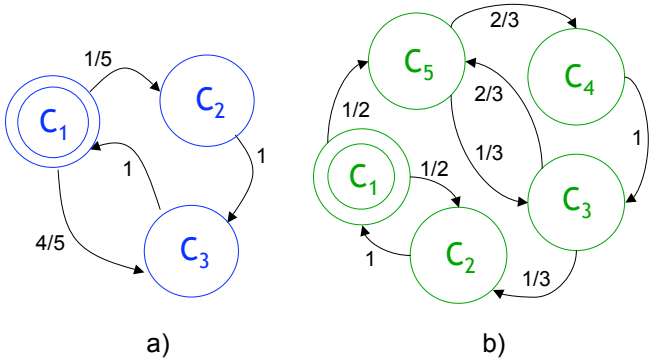

Fig. 4. Examples of Markov chains with a) low and b) high entropy rate.

Considering Equation 2, it can be noticed that the value of the entropy inherently takes into account the number of concepts present in each story unit (*i.e.,* the number of states N) and the grade of interconnections between concepts (by the transition probability distribution Δ). Markov entropy rate $H(X)$ can thus be considered a compact but effective index able to describe the organization and structure informativeness of a story unit, and therefore it can be used for retrieval purposes and scene classification.

6. RETRIEVAL PERFORMANCES

In order to objectively evaluate the effectiveness of the Markov entropy index to retrieve similar story unit in terms of structure informativeness, we carried out some experiments using seven videos (see Table 1). In particular one news programme and six feature movies have been investigated, for a total time of about ten hours of video and several thousands of shots.

Table 1. Video data set.

No.	Video (genre)	hh:min:sec	# of *LSU*
1	*Tg La7 (news)*	00:36:53	20
2	*Johnny Stecchino (movie)*	01:55:00	37
3	*Kill Bill II (movie)*	02:11:18	61
4	*A Beautiful Mind (movie)*	00:20:30	15
5	*Notting Hill (movie)*	00:39:53	53
6	*Man on the Moon (movie)*	01:53:42	51
7	*Matrix (movie)*	02:10:45	80

A software prototype for story unit retrieval has been realised based on the preceding discussion; the test set comprises a total of more than three hundred story units which present different grades of structure informativeness and complexity. For the sake of experimental evaluation, stories have been manually annotated into three general categories as in [15], that are "dialogue", "progressive" and "hybrid" stories (see Table 2).

Table 2. Story units belonging to different categories.

***LSU* category**	**# of *LSU***
Dialogue	91
Progressive	107
Hybrid	119
Total	**317**

The retrieval system adopts a query-by-example paradigm: for each query provided by the user, the first 30 most similar story units (in terms of structure information) are retrieved according to L^1 metric. Examples of queries belonging to different *LSU* categories, *i.e.,* "dialogue", "progressive" and "hybrid" are displayed in Figures 5, 6 and 7, respectively.

Fig. 5. Example of a query story unit: "dialogue".

Fig. 6. Example of a query story unit: "progressive".

Fig. 7. Example of a query story unit: "hybrid".

Fifty queries by example have been performed for each category ("dialogue", "progressive" and "hybrid") by randomly choosing story units as inputs, and for each query, best 30 retrieved results are presented in a ranked list to the user.

Precision results on *LSU* category in Table 3 are obtained by considering the best 30 retrieved stories and by averaging on fifty trials. Best results are obtained on "dialogues". Since only 30 retrievals are presented to the user, for this search prototype recall has not been considered.

Results may be further improved by removing minor errors occurring on story boundaries, since the process is completely automated (from shot boundary detection to Markov entropy computation). Note that, for this retrieval application, only story units with $H(X) \neq 0$ (*i.e.,* not completely predictable) have been considered.

Table 3. Average retrieval accuracy on first 30 retrieved story units.

	Retrieved *LSU* (%)		
Query *LSU* category	*Dialogue*	*Hybrid*	*Progressive*
Dialogue	**73.30 %**	10.00%	16.70%
Hybrid	30.00%	**56.67 %**	13.33%
Progressive	6.67%	33.33%	**60.00 %**

7. CONCLUSIONS

In this paper we have proposed a method to retrieve story units which present similar structure given a query one. The index used to retrieve similar units is the Markov entropy computed on a Markov chain used to model the story units. Such index takes into account the number of expressed visual concepts and the pattern in which they appear inside the story. The effectiveness of the proposed approach is demonstrated on different videos, and encouraging results have been obtained on retrieved results. Future work aims to integrate this search method, which is based only on story structure, with other effective descriptors describing visual and audio story content and pace.

8. ACKNOLEDGMENTS

This research work has been partially supported by EU project *RUSHES (FP6-045189)*. We would also like to thank Dr. Nicola Adami and Mr. Roberto Ceretti for fruitful discussions and precious help during the experimental evaluation.

9. REFERENCES

[1] 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.

[2] Q. Huang, Z. Lou, A. Rosenberg, D. Gibbon, and B. Shahraray, "Automated generation of news content hierarchy by integrating audio, video, and text information," in *Proc. of ICASSP '99*, March 1999, vol. 6, pp. 3025–3028.

[3] Z. Xiong, P. Radhakrishnan, and A. Divarakan, "Generation of sports highlights using motion activity in combination with a common audio feature extraction framework," in *Proc. ICIP'03*. Barcelona, Spain, Sept. 2003.

[4] A. Hanjalic, "Extracting moods from pictures and sounds," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, March 2006.

[5] 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, June 1999.

[6] J. Vendrig and M. Worring, "Systematic evaluation of logical story unit segmentation," *IEEE Trans. on Multimedia*, vol. 4, no. 4, pp. 492–499, December 2002.

[7] 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.

[8] 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.

[9] A.A. Alatan, A.N. Akansu, and W. Wolf, "Comparative analysis of hidden markov models for multi-modal dialogue scene indexing," in *Proc. of ICASSP'00*. Istanbul, Turkey, 6-8 June 2000.

[10] S. Benini, P. Migliorati, and R. Leonardi, "A statistical framework for video skimming based on logical story units and motion activity," in *Proc. of CBMI'07*. Bordeaux, France, 25-27 June 2007.

[11] G. Cees, M. Snoek, and M. Worring, "Multimodal video indexing: A review of the state of art," *Multimedia tools and application*, vol. 25, no. 1, pp. 5–35, 2005.

[12] S. Benini, A. Bianchetti, R. Leonardi, and P. Migliorati, "Extraction of significant video summaries by dendrogram analysis," in *Proc. of ICIP'06*. Atlanta, GA, USA, 8-11 Oct 2006.

[13] 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.

[14] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley-Interscience, New York, NY, USA, 1991.

[15] H. Sundaram and S.-F. Chang, "Determining computable scenes in films and their structures using audio-visual memory models," in *Proc. of ACM*. Los Angeles, CA, USA, November 2000, pp. 95–104.