RETRIEVAL OF VIDEO SCENES BY STRUCTURAL DESCRIPTORS

Ugo Ciracì[1], *Luca Buriano*[2] *and Riccardo Leonardi*[1]

[1]University of Brescia, DEA, v. Branze 38, Brescia, Italy
[2] Telecom Italia Lab, v. Reiss Romoli 274, Torino, Italy

Email: ugo.ciraci@ing.unibs.it, luca.buriano@telecomitalia.it, riccardo.leonardi@ing.unibs.it

ABSTRACT

Structural information, on the form of shot patterns and timing, plays an important role in scripted video, being consistently used by filmmakers as a way to convey meaning and significance into their work. For this reason, a video retrieval system can benefit from exploiting this kind of information. In this paper we propose a novel method for the retrieval of video scenes according to their structural similarity, based on graph-theoretical measures and vector quantization techniques. We show the results of our method in some scene retrieval experiments on a data set of 802 movie scenes, extracted from a set of 180 mainstream movies.

1. INTRODUCTION

The development of flexible video retrieval and browsing applications asks for the availability of new methods for the automatic extraction of similarity measures between video segments.

Structural similarity between video segments deals with the features that can be abstracted from the specific audiovisual content, such as shot occurrence patterns and shot timing. This is especially relevant in the case of movies (and more generally in the case of scripted video), because filmmakers make use of montage as an essential part of their artistic creative process, organizing the shots according to different structural patterns in order to convey semantics, significance and style into movie scenes.

In this paper we propose a novel method for the automatic extraction of structural descriptors from video scenes, the estimation of the similarity between them and, given a scene as a query, the retrieval of similar scenes from a structural point of view.

The paper is organized as follows. In §2, the structural representation of video scenes used in this work is introduced. In §3 and §4 the core of our method, i.e. the structural descriptors and the similarity measures between them, is explained. In §5 we show the results of some video scene retrieval experiments on real video data; finally, in §6 and §7, the relationships of our method with other works and its possible future evolutions are discussed.

2. STRUCTURAL REPRESENTATION

At the simplest structural level, a video segment can be seen as a sequence $\mathbf{S = S_1, S_2,\ldots,S_n}$ of symbols denoting the shots belonging to the scene. When the shots in the sequence share a common semantic thread, the sequence is called a *Logical Story Unit* (LSU) [1]. The shots in a LSU can be grouped into different clusters, each representing a common concept; for instance, in a dialog alternating the views of different speakers, the shots corresponding to the views of the same speaker will belong to the same cluster. From the computational point of view, the grouping of the shots can be obtained clustering them on the basis of visual similarity measures [2]. This clustering process leads to a representation of the LSU as a sequence $\mathbf{C = C_1, \ldots, C_m}$ of shot clusters. For instance, given the sequence of shots **{S1, S2, S3, S4, S5, S6, S7}** and the clusters **C1 = {S1, S3, S5}, C2 = {S2, S4, S6}**, **C3 = {S7}**, the corresponding shot cluster sequence will be **C = C1, C2, C1, C2, C1, C2, C3.** Finally, the shot cluster sequence can be mapped on a weighted directed graph **G,** called *scene transition graph (STG)* [3]**,** having the clusters as nodes and the transitions between the clusters in the sequence as arcs. In the present work, the weight of an arc connecting two nodes is given by the number of transitions occurring in the sequence between the two corresponding clusters.

Besides, each shot in a LSU has a given time duration, usually called *shot length.* The *shot length sequence* $\mathbf{SL = SL_1,\ldots,SL_n}$ of single shot lengths from a LSU gives a representation of its temporal features, such as the pace of the action in the scene.

In this work, we will use the scene transition graph **G** and the shot length sequence **SL** as a representation of the structural features of a given LSU.

3. STRUCTURAL DESCRIPTORS

Structural descriptors are extracted from the representation explained in the previous paragraph, in order to encode relevant structural features of LSUs in a more compact and uniform way.

3.1. Scene Transition Graph descriptor

Information about shot occurrence patterns is captured by analyzing the topology of the scene transition graph, according to the following steps:

1. For each node in the scene transition graph **G**, a set of graph-theoretic measures are extracted:
    - the *indegree* of the node, i.e. the number of arcs leading to the node;
    - the *outdegree* of the node, i.e. the number of arcs leading away from the node;
    - the *betweenness centrality* of the node [4], a measure of the centrality of the node in the graph, taking into account the weights of the directed graph's arcs;
    - the *characteristic path length* of the node [4], measuring the average distance between that node and any other node in the graph, taking into account the weights of the arcs.

    The result of this step is a set of graph measure vectors [*indegree, outdegree, betweenness, pathlength]*, one vector for each node in **G**.
2. A Vector Quantization (VQ) [5] algorithm is applied to the set of graph measure vectors obtained in the previous step, clustering this set into **k** representative vectors (codewords), called in this paper *node codewords*. The VQ algorithm used determines automatically the number of clusters **k** that provides the best approximation, in a range from 1 to a maximum given value $\mathbf{max_k}$ (in this work, $\mathbf{max_k}$ =3).

    The result of this step is a Scene Transition Graph structural descriptor **STGD = {NC, NS}** where **NC** is the *node codebook*, i.e. the set of the **k** node codewords, and **NS** is the *node signature,* i.e. the vector of **k** components, giving the percentage of the nodes in the graph **G** belonging to each node codeword.

3.2. Shot length descriptor

Temporal information is taken into account by analyzing the distribution of shot length values, according to the following steps:

1. The shot length sequence **SL** is aggregated according to the nodes in **G**, each node corresponding to one or more shots. For each node in **G,** the shot length average (SLA) and shot length standard deviation (SLSD) for that node are computed.

    The result of this step is a set of vectors **[*SLA, SLSD*]**, one vector for each node in **G**.
2. As for the computation of the Shot Transition Graph descriptor, a Vector Quantization algorithm is applied to the set of vectors obtained in the previous step.

    The result of this step is a Shot Length structural descriptor **SLD = {LC, LS}** where **LC** is the *shot length codebook*, i.e. the set of shot length codewords, and **LS** is the corresponding *shot length signature*.

The time complexity of the descriptor extraction algorithm is dominated, in this work, by the complexity of the weighted betweenness centrality computation, that is $O(\mathbf{ma}+\mathbf{m}^2\log \mathbf{m})$ (using Brandes' fast algorithm [6]), where **m** is the number of nodes and **a** the number of arcs in the scene transition graph.

4. SIMILARITY ESTIMATION

The distance (conversely, the similarity) between two LSUs from the structural point of view is computed as a pair of two separate distances $\{\mathbf{d_{STGD}}, \mathbf{d_{SLD}}\}$ for the corresponding STG descriptors and SL descriptors:

- The distance between two STG descriptors $\mathbf{STGD_a} = \{\mathbf{NC_a}, \mathbf{NS_a}\}$ and $\mathbf{STGD_b} = \{\mathbf{NC_b}, \mathbf{NS_b}\}$ is computed using the *Earth Mover's Distance (EMD)* [7]*:*

    $$\mathbf{d_{STGD} = EMD\ (NS_a, NS_b, NCM_{ab})}$$

    where $\mathbf{NCM_{ab}}$ is the cost matrix associated with the node codebooks $\mathbf{NC_a}$ and $\mathbf{NC_b}$**.**
- In a similar way, the distance between two SL descriptors $\mathbf{SLD_a} = \{\mathbf{LC_a}, \mathbf{SLS_a}\}$ and $\mathbf{SLD_b} = \{\mathbf{LC_b}, \mathbf{LS_b}\}$ is computed using the EMD*:*

    $$\mathbf{d_{SLD} = EMD\ (LS_a, LS_b, LCM_{ab})}$$

    where $\mathbf{LCM_{ab}}$ is the cost matrix associated with the shot length codebooks $\mathbf{LC_a}$ and $\mathbf{LC_b}$**.**

In this work, the composition of the two distances $\{\mathbf{d_{STGD}}, \mathbf{d_{SLD}}\}$ is performed using the following *order-based rank aggregation* method [8] [9], a general technique for mixing the rankings induced by different distances in a retrieval task.

Given a query item **X**, a set of items $\{\mathbf{Y_1},\ldots,\mathbf{Y_n}\}$ and a set of distance functions $\mathbf{d_1(X,Y),..,d_z(X,Y)}$ defined on these items, let $\mathbf{R_i(Y_j)}$ be the ranking of the item $\mathbf{Y_j}$ obtained applying the distance function $\mathbf{d_i}$ to all the pairs $\mathbf{(X, Y_j)}$ and sorting the resulting distance values in ascending order. Then for a given item $\mathbf{Y_j}$ the aggregated rank $\mathbf{AR(Y_j)}$ is given by:

$$\mathbf{AR(Y_j) = \Sigma(R_1(Y_j),\ldots,R_z(Y_j))}$$

i.e. the sum of the rankings of $\mathbf{Y_j}$ with respect to the different distance functions.

5. EVALUATION

In order to assess the suitability of our method for structural similarity estimation and scene retrieval, some experiments have been performed on real video data:

- The video stream of a set of 180 mainstream movies is automatically segmented in LSUs. Each LSU is given the structural representation explained in §2, using the methods described in [2] and [10]. The LSUs' timescale is of tens of seconds to minutes (and of tens of shots), giving scene transition graphs with sizes ranging from a few nodes to a few tens of nodes.
- From the global set of LSUs, a data set of 802 LSUs is randomly selected, computing for each LSU the structural descriptors described in §3. Graph-theoretical measures are calculated using the Brain Connectivity Toolbox by Olaf Sporns [11].
- Each LSU in the data set is manually annotated as belonging to one or more of the following scene categories:
    - *Dialogue 2*: a scene with a dialogue between two characters;
    - *Dialogue 3+*: a scene with a dialogue among three or more characters;
    - *Talking*: a generic talking scene (including Dialogue 2, Dialogue 3+, and other talking situations);
    - *Action*: an action scene;
    - *Fight/War* : a fighting or war scene;
    - *Slow progression*: a scene with a (quite slow) movement/progression from a place/situation to another, or a "parallel montage" scene between two different places/situations.

  Other categories are of course possible, the ones above being chosen mainly because they are likely to show structural diversity and are suitable for a quick manual annotation, allowing getting a quick proof-of-concept of the method.
- Following a Query by Example (QBE) paradigm, each LSU in the data set is taken, one at a time, as the current query LSU. The distance between the query LSU and all the other LSUs in the data set is computed as explained in §4; distance values are sorted in ascending order, ranking the LSUs in the data set from the most similar to the query LSU to the least similar.
- Retrieval precision for a given query LSU, belonging to a given category, is calculated as the fraction of the LSUs appearing as the first **r** results that share the same category of the query LSU. Because a LSU can belong to more than one category, precision is calculated separately for different categories.

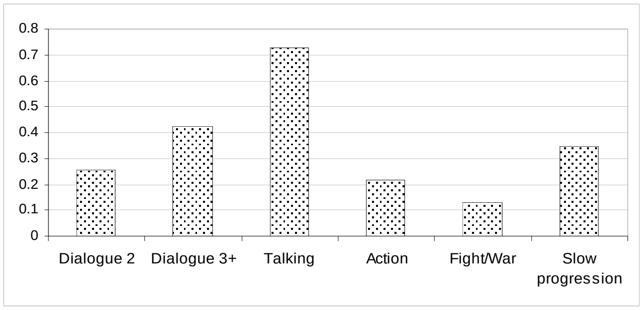

Figure 1

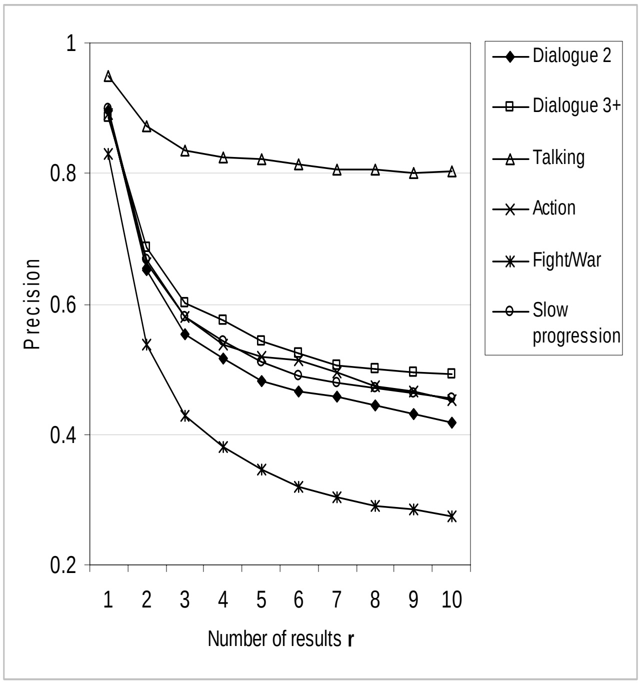

Figure 2

Figure 1 shows the fraction of LSUs belonging to each category with respect to the number of LSUs in the whole data set. Figure 2 summarizes the precision values obtained for different categories and for different values of **r,** averaged over the entire data set.

6. RELATED WORK

Due to the growing interest in video retrieval systems, in recent years many methods for video scene similarity, retrieval and classification have been proposed, mainly focusing on the use of audiovisual features extracted from the audio/video stream. Examples include methods for motion-based shot retrieval [9], face detection-based shot retrieval [12], movie scene retrieval [13], sports video scene analysis [14][15], video genre classification [16], video news event detection [17] and affective video content analysis [18].

Benini et al. [10] propose a structural similarity and retrieval framework based on Markov entropy measures. Their method is somewhat complementary to ours, giving a powerful way of classifying LSUs into broad categories by a single entropy value, while our methods aims at capturing more specific features of scene graphs with a more complex descriptor framework, taking also into account temporal features. Therefore, the combination of the two methods could lead to interesting results.

7. FUTURE WORK

The evolution of the work presented in this paper will follow three main directions.

First, the combination of different graph-theoretical measures, vector aggregation strategies, distance measures and distance compositions will be explored, in order to improve the overall performance of the method as well as its performance with respect to a specific category set.

Next, the proposed method is not intended to work as a standalone retrieval system: future research will aim to integrate it as a building block of a more complete video retrieval framework, exploiting both structural and audiovisual features.

Finally, although the method presented in this paper has been developed to address a video retrieval task, we believe that the core of the method itself is general enough to be adapted and extended to similarity and retrieval tasks in other areas where graph-theoretical representations play an important role, such as the analysis of social network, biochemical pathway and neural connectivity patterns.

8. REFERENCES

[1] 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.

[2] 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.

[3] M. M. Yeung and B.-L. Yeo, “Time-constrained clustering or segmentation of video into story units,” in Proc. of ICPR’96. Vienna, Austria, Aug 1996, vol. III - vol.7276, p. 375.

[4] O. Sporns, Graph theory methods for the analysis of neural connectivity patterns. Kötter, R. (ed.) Neuroscience Databases. A Practical Guide. Klüwer (2002).

[5] A. Gersho and R. M. Gray, “Vector Quantization and Signal Compression”, Kluwer Academic Press/Springer, 1992.

[6] U. Brandes, “A faster algorithm for betweenness centrality”, J Math Sociol 2001; 25: 163-177.

[7] Y. Rubner; C. Tomasi, L. J. Guibas, "A Metric for Distributions with Applications to Image Databases", in IEEE International Conference on Computer Vision, pages 59-66, January 1998.

[8] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, “Rank Aggregation Methods for the Web”, in Proceedings of the 10th International World Wide Web Conference. 2001, 613-622.

[9] S. Benini, L.-Q. Xu, R. Leonardi, "Using lateral ranking for motion-based video shot retrieval and dynamic content characterization," in Proceedings of CBMI'05, Riga, Latvia, June 21-23, 2005.

[10] S. Benini, A. Bianchetti, R. Leonardi, and P. Migliorati, “Retrieval of Video Story Units by Markov Entropy Rate”, CBMI 2008.

[11] http://www.indiana.edu/%7Ecortex/connectivity.html

[12] J. Sivic, M. Everingham, and A. Zisserman. “Person spotting: video shot retrieval for face sets”, in International Conference on Image and Video Retrieval (CIVR), pages 226–236, 2005.

[13] Y. Hun-Woo, “Retrieval of movie scenes by semantic matrix and automatic feature weight update”, in Expert Systems with applications: An International Journal, Volume 34, Issue 4 (May 2008).

[14] 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, 2003.

[15] 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.

[16] 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.

[17] D. Xu, S.-F. Chang, “Video Event Recognition Using Kernel Methods with Multilevel Temporal Alignment”, IEEE Transactions on Pattern Analysis and machine Intelligence, vol. 30, no. 11, Nov. 2008.

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