A Video Summarization Method Based on Spectral Clustering

Marcos Vinicius Mussel Cirne and Helio Pedrini

Institute of Computing - University of Campinas
Campinas, SP, Brazil, 13083-852

Abstract. The constant increase in the availability of digital videos has demanded the development of techniques capable of managing these data in a faster and more efficient way, especially concerning the content analysis. One of the research areas that have recently evolved significantly at this point is video summarization, which consists of generating a short version of a certain video, such that the users can grasp the central message transmitted by the original video. Many of the video summarization approaches make use of clustering algorithms, with the goal of extracting the most important frames of the videos to compose the final summary. However, special clustering algorithms based on a spectral approach have obtained superior results than those obtained with classical clustering algorithms, not only in video summarization techniques but also in other fields, such as machine learning, pattern recognition, and data mining. This work proposes a method for summarization of videos, regardless of their genre, using spectral clustering algorithms. Possibilities of algorithm parallelization for the purpose of optimizing the general performance of the proposed methodology are also discussed.

1 Introduction

Due to the great increase in the generation of digital videos in the last years, there is an increasingly need to develop techniques that are capable of manipulating these data in an automatic, efficient and accurate way, concerning the issues of searching, browsing, retrieval and content analysis. Among these techniques is the video summarization, which consists of deriving a short version from a given video, preserving as much relevant information as possible, such that the users can grasp the message transmitted by the original video. The generated summaries can then be integrated into many applications, such as interactive searching and browsing systems, making both management and access to video content more accurate [14].

Nevertheless, defining what is important or not in video summarization is an open problem, especially because there is a variety of video genres, such as sports, movies, news programs, documentaries, and home movies in general. Even to humans, it is hard to reach a consensus to know how good a summary is, since what is relevant to ones may not be to others. Thus, the main challenge of the video summarization field is in how to make a system to take the best decisions to choose the most important parts of a video. This is usually done by analyzing

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 479–486, 2013.

high level features (e.g. semantic content, time, space) or low level features (e.g. color histograms, texture, subtitles, audio, shape and motion descriptors).

Among the various approaches to the video summarization problem are those which make use of clustering algorithms, that are also objects of study in fields such as data mining, machine learning, and statistics. The idea beyond these approaches is to split the frames of a given video into different groups such that frames that belong to the same group are more similar among themselves. Then, a set of keyframes is extracted from these groups, i.e., the frames that best represent both the belonging groups and the essence of the original video content. Later, the final summary is generated from these keyframes.

A clustering technique that has been increasingly growing recently is the spectral clustering [13], due to the fact that it can generate more satisfactory results than those obtained by classic clustering algorithms. In the case of video summarization, even though there are many approaches that use clustering algorithms, little has been produced with spectral clustering algorithms so far.

The objective of this work is to propose and analyze a new method for video summarization of any genre using spectral clustering algorithms. A qualitative analysis of the generated summaries with different feature descriptors is conducted, comparing the results with a specific database, which includes summaries from other approaches.

The main contributions of this work include the creation and implementation of a method that can be integrated into many video processing environments and a performance and accuracy analysis of the proposed method, considering the variety of existing video genres.

This paper is organized as follows: Section 2 describes the main concepts about video summarization and spectral clustering, as well as works related to both topics; Section 3 defines the proposed methodology for this work; Section 4 presents and discusses some of the obtained results with the proposed method; Section 5 includes the general conclusions about the discussed topic and some future work suggestions in order to improve the proposed method.

2 Concepts and Related Work

This section describes general concepts about video summarization, together with related works and spectral clustering algorithms, and how they are applied to the video summarization context.

2.1 Video Summarization

A digital video can be defined as a collection of images that have the same dimensions, grouped according to a temporal sequence. Each of these images is known as *frame*, which corresponds to the smallest structural unit of a video, representing a picture captured by a camera in a given time instant of the video. The frames can be grouped into *shots*, which are sequences of frames, captured in a contiguous way, and that represent a continuous action in time or space. Finally, a group of shots that are semantically correlated constitutes a *scene*.

Video summarization techniques can be divided into *static* and *dynamic*. In the first category, the summary is generated as a collection of still images denominated *keyframes* [16], that represent the content of a video in the form of a storyboard. The advantage of this approach is in its simplicity and efficiency, usually being free of redundancies, but it may not preserve the temporal order of the selected keyframes. In the second category, many segments of the video are chosen, which are then organized such that the temporal order of the video is preserved [21]. Dynamic summarization has the main advantage of generating summaries which a higher richness of details, but it is more expensive than static summarization approaches, besides the possible generation of redundancies.

Another challenge in the video summarization field is the definition of standard metrics to evaluate the quality of the results. At the moment, there is no consistent platform to evaluate summaries. Thus, each work has its own evaluation method and, in most cases, it does not compare the results with other existing methods [20].

2.2 Spectral Clustering

Spectral clustering [13] has become one of the most popular clustering techniques lately, being an important research object in fields such as pattern recognition, machine learning, and signal processing. It provides better results than those from classic clustering algorithms (such as K-means) and it can be easily implemented by means of numeric computation platforms. In the video summarization context, spectral clustering can be used in tasks such as keyframe extraction [7] and shot boundary detection [8].

Given a set of n points, located at an l-dimensional space, to be divided into k distinct subsets, where n, l and k are positive integers, an affinity matrix $A_{n\times n}$ is constructed such that each element $A(i,j)$ corresponds to a similarity measure $s_{ij} \geq 0$ that represents the likelihood degree between a pair of points i and j of the set, with $A(i,i) = 0$. Thus, the bigger the value of $A(i,j)$, the higher is the similarity between the points i and j and vice-versa.

Later, the diagonal matrix $D_{n\times n}$ is defined, where $D(i,i) = \sum_{j=1}^{n} A(i,j)$. From A and D, the Laplacian matrix $L = I - (D^{-1/2}AD^{-1/2})$ is constructed, where $I_{n\times n}$ is the identity matrix. In the next step, the k largest eigenvectors of L are calculated, forming the matrix $X = [x_1 x_2 ... x_k]$ by stacking these eigenvectors in k columns. After that, the matrix Y is created from X by normalizing the rows of X such that each one has unitary length. Finally, the rows of Y are separated into k groups by the K-means algorithm (or any other clustering algorithm, such as the ones described in [9]), assigning the point i of the initial set to group j if, and only if, the row i of matrix Y is assigned to cluster j.

The choices of the similarity measure to be used and the number of clusters in which the dataset is split are not trivial tasks, once that they are subject to the application domain of this set. First of all, it must be assured that the data considered as "very similar" by the chosen similarity measure have a very close relationship in the application domain as well [13]. Furthermore, in most of the

cases, there is not a "correct" number of groups. In this situation, it is common to use strategies that find this number in an automatic way [18].

Usually, the matrices computed by the spectral clustering algorithms are very large, demanding a large storage space, especially when working with digital videos, composed of a considerable number of frames. In order to guarantee the efficiency on the implementation of these algorithms, it is necessary that the Laplacian matrix related to the similarity graph be sparse, simplifying the task of calculating the k largest eigenvectors. To do this, graphs such as ϵ-neighborhood and k-nearest neighbors are used, eliminating the computation of the similarity measures between every single pair of points.

3 Methodology

The methodology of this work will be focused on a new method for digital video summarization of any genre using spectral clustering to obtain summaries with a better quality than those found in the state-of-the-art. A comparative analysis of the generated summaries of some methods of the literature with the ones generated by the proposed method is conducted. A general flowchart of the methodology stages is shown in Figure 1.

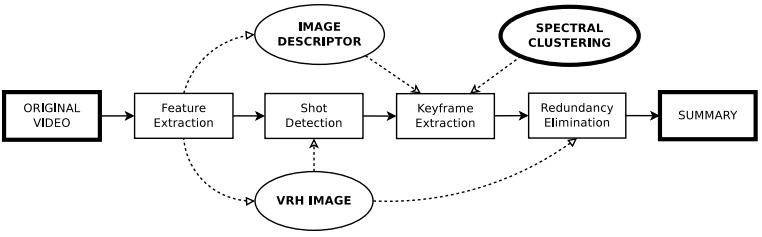

Fig. 1. Flowchart of the main stages of the proposed method for video summarization

From a given digital video, the feature extraction stage will primarily make a sampling of this video in frames. To optimize the performance of the application, only 5 frames per second are used in this stage. From these frames, both the visual rhythm by histogram (VRH) [11] and image descriptors for each frame are calculated. In this process, many image descriptors that encompass spatial and temporal features are evaluated, such as SIFT (Scale-Invariant Feature Transform) [12], SURF (Speeded Up Robust Features) [5] and ORB (Oriented FAST and Rotated BRIEF[1]) [17].

In the shot detection stage, the estimation of the number of video shots is started, which will be the number k of clusters used in the next stage. From the VRH image, the shot boundaries are detected by using the local adaptive threshold technique described in [19], which produces more accurate results rather than

[1] Acronym that stands for Binary Robust Independent Elementary Features [6].

using a fixed threshold to detect the boundaries. Starting with $k = 1$, every time a shot boundary is detected, k is incremented by 1.

After estimating k, a spectral clustering algorithm is executed for the keyframe extraction stage. Using descriptor feature vectors, an affinity matrix A is constructed, as defined in Section 2.2, where the element $A(i, j)$ corresponds to the distance between the feature vectors of frame i and the one of frame j. After the calculation of the normalized eigenvectors, the K-means algorithm is run to cluster the frames according to the shots to which they are associated, where the number of clusters corresponds to k.

Finally, the keyframes of each cluster are extracted based on the centroids calculated by K-means (one keyframe per cluster) and preserving their temporal order. The selected keyframes correspond to the ones that are closest to their respective cluster centroids. Before the summary generation process, a post-processing is performed to eliminate redundant frames. This is done by computing the sums of pairwise pixel distances between the columns of the VRH image (generated in the feature extraction stage) related to two consecutive keyframes. After that, these values are compared to a distance threshold T_d. If the distance between keyframe i and keyframe $i + 1$, where $1 \leq i \leq k - 1$, is less than T_d, the keyframe i will be considered as redundant and, therefore, will not be included in the final summary. The threshold value was empirically defined as $T_d = (\mu_d + \sigma_d)/4$, where μ_d and σ_d are the mean and the standard deviation of all distances, respectively. This approach performs well with most of the generated redundant frames from the videos used in the tests, but it may fail at detecting redundant frames with high luminosity differences (brightness and contrast), since their columns in the VRH image are very distant from each other.

From the remaining keyframes, the final summary is then created, which can be done in a static way, generating a storyboard, or in a dynamic way, taking a certain amount of frames around each keyframe in the original video, such that the total number of selected keyframes correspond to a percentage of the total number of frames of the original video.

The advantage of this method is that every stage is executed in an unsupervised way, such that the number of shots does not need to be known *a priori*. However, the whole summarization process is still expensive, because of the spectral clustering, even though it leads to more accurate results than standard clustering approaches.

4 Experimental Results

The tests were done using an AMD Phenom II X6 3.2 GHz processor and 4 GB of memory. The methodology described in Section 3 was implemented with OpenCV platform [1]. A collection of 50 videos of several genres from Open Video Project (OVP) [2], available at the VSUMM database [3] (provided by the authors of the approach described in [4]), were used in the tests, together with the respective summaries produced by different video summarization methods, which include Delaunay Triangulation (DT) [15], STIMO (STIll and MOving Video Storyboards) [10],

as well as the OVP summaries and the ones provided by VSUMM. All of the videos have, together, approximately a total duration of 75 minutes and 150,000 frames (352 × 240 pixels). After the execution of the implementation of the proposed method for each descriptor and using all videos, it was observed that SIFT provided the fastest execution time, with a total execution time of 1.10 hours, followed by ORB (4.04 hours) and SURF (7.59 hours).

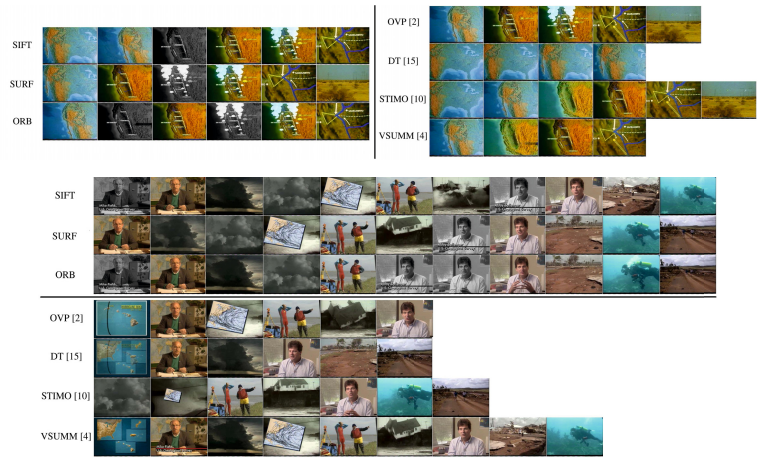

Fig. 2. Summarization results for *The Great Web of Water, Segment 02*) video (upper image) and *Hurricane Force - A Coastal Perspective, Segment 03* (lower image). For each descriptor, redundant frames are represented as greyscale images.

To evaluate the quality of the summaries, only two videos are analyzed due to space limitation in the paper: *The Great Web of Water, Segment 02*, which has 5 shots, and *Hurricane Force - A Coastal Perspective, Segment 03*, with 12 shots. Figure 2 shows the respective results, together with the summaries generated by different approaches, as well as the one provided by the OVP database. For the first video, it can be seen that the proposed method generated summaries with 6 keyframes, one more than the number of shots, which means that the shot boundary detection process performed very well for this video. Also, the redundant frames (represented as greyscale images) were properly detected and eliminated for the final summary, once that the respective contents of the detected redundant frames are similar to their consecutive frames, leaving only the colored ones. With respect to the quality of the summaries, the SURF summary was the only descriptor that included the contents of all shots, being the closest to the OVP summary. Furthermore, the SIFT summary included two keyframes of a same shot (1st and 2nd frames), and the ORB summary was the one that generated more redundant frames (2nd and 4th frames) than the other descriptors. Comparing to other approaches, SURF performed slightly better than both STIMO and VSUMM, which produced the best summaries among the other approaches. This happens because STIMO included more than one frame of a shot,

even though it included at least one frame of every shot, and VSUMM missed the last shot.

For the second video, all of the summaries of each descriptor contain 11 keyframes, one less than the number of video shots. In the redundancy elimination process, it can be noticed that three frames were discarded in the ORB summary (1st, 6th and 7th frames), whereas SIFT summary had two discarded frames (1st and 8th frames) and only one for the SURF summary (7th frame). However, all of the eliminated frames (except for the 7th one of the ORB summary) have a little more information than the remaining consecutive frames of the respective final summaries. Concerning the summary content, SURF selected most of the different shots not only among the descriptors but also the other approaches as well. On the other hand, comparing the summaries of the proposed method to the OVP summary, none of them was able to select a frame from the first shot, as occurred both in DT and VSUMM summaries.

Despite this analysis, it is hard to evaluate how the misdetection of a shot (i.e., when a frame of a shot is not included in the final summary) affects the comprehension of the central message transmitted by a video. For that, a more subjective evaluation must be made, once it requires a deeper content analysis and a general consensus about the degree of relevance of each shot. In other words, even though the summaries produced by each descriptor have more different shots than the ones of other approaches (including the OVP), all of them may have the same relevance in particular situations.

5 Conclusions

This paper described a method for video summarization from any genre using a spectral clustering algorithm. Different image descriptors were used to extract features from the video frames, as well as the normalized eigenvectors of the respective affinity matrices. The K-means algorithm was used to cluster video frames according to the number of shots detected by a previous procedure that uses a visual rhythm by histogram image to identify shot boundaries. Redundant frames are then discarded to produce the final summaries, which were compared against summaries produced by different video summarization approaches (DT, STIMO, VSUMM and the ground-truth provided by the OVP database).

Despite the slowest processing time, the summaries produced by SURF were the best among the tested descriptors, once they detected most of the different shots and generated less redundant frames than SIFT and ORB. Comparing SURF to other approaches, the results were very close in most cases, although SURF produced more complete summaries. Furthermore, both the shot boundary detection and the redundancy elimination procedures performed well in the analyzed videos, yet they still need some adjustments to improve their accuracy.

Acknowledgements. The authors are grateful to FAPESP and CNPq for the financial support.

References

1. OpenCV: Open Source Computer Vision (2013), `http://opencv.org`

2. The Open Video Project (2013), http://www.open-video.org
3. VSUMM (Video SUMMarization) (2013), https://sites.google.com/site/vsummsite
4. de Avila, S.E.F., Lopes, A.P.B., da Luz Jr., A., de Albuquerque Araújo, A.: VSUMM: A Mechanism Designed to Produce Static Video Summaries and a Novel Evaluation Method. Pattern Recognition Letters 32(1), 56–68 (2011)
5. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
6. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: Binary Robust Independent Elementary Features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 778–792. Springer, Heidelberg (2010)
7. Chasanis, V., Likas, A., Galatsanos, N.: Video Rushes Summarization Using Spectral Clustering and Sequence Alignment. In: 2nd ACM TRECVid Video Summarization Workshop, Vancouver, BC, Canada, pp. 75–79 (2008)
8. Damnjanovic, U., Izquierdo, E., Grzegorzek, M.: Shot Boundary Detection Using Spectral Clustering. In: 15th European Signal Processing Conference, Poznan, Poland, pp. 1779–1783 (September 2007)
9. Elhamifar, E., Sapiro, G., Vidal, R.: See All by Looking at a Few: Sparse Modeling for Finding Representative Objects. In: IEEE Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, pp. 1600–1607 (2012)
10. Furini, M., Geraci, F., Montangero, M., Pellegrini, M.: STIMO: STIll and MOving Video Storyboard For The Web Scenario. In: Multimedia Tools and Applications, vol. 46, pp. 47–69. Kluwer Academic Publishers, Hingham (2010)
11. Guimarães, S.J.F., Couprie, M., Araújo, A.D.A., Leite, N.J.: Video Segmentation Based on 2D Image Analysis. Pattern Recognition Letters 24(7), 947–957 (2003)
12. Lowe, D.: Object Recognition from Local Scale-Invariant Features. In: Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
13. Luxburg, U.: A Tutorial on Spectral Clustering. Statistics and Computing 17(4), 395–416 (2007)
14. Money, A.G., Agius, H.: Video Summarisation: A Conceptual Framework and Survey of the State of the Art. Journal of Visual Communication and Image Representation 19(2), 121–143 (2008)
15. Mundur, P., Rao, Y., Yesha, Y.: Keyframe-Based Video Summarization Using Delaunay Clustering. International Journal on Digital Libraries 6, 219–232 (2006)
16. Peng, J., Xiaolin, Q.: Keyframe-Based Video Summary Using Visual Attention Clues. IEEE MultiMedia 17(2), 64–73 (2010)
17. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: An Efficient Alternative to SIFT or SURF. In: IEEE International Conference on Computer Vision, Barcelona, Spain (2011)
18. Sanguinetti, G., Laidler, J., Lawrence, N.D.: Automatic Determination of the Number of Clusters Using Spectral Algorithms. In: IEEE Machine Learning for Signal Processing, pp. 28–30 (2005)
19. Shekar, B., Raghurama Holla, K., Sharmila Kumari, M.: Video Shot Detection Using Cumulative Colour Histogram. In: Mohan, S., Kumar, S.S. (eds.) 4th International Conference on Signal and Image Processing. LNEE, vol. 222, pp. 353–363. Springer, Heidelberg (2012)
20. Truong, B.T., Venkatesh, S.: Video Abstraction: A Systematic Review and Classification. ACM Transactions on Multimedia Computing, Communications and Applications 3(1) (February 2007)
21. Zhou, H., Sadka, A.H., Swash, M.R., Azizi, J., Sadiq, U.A.: Feature Extraction and Clustering for Dynamic Video Summarisation. Neurocomputing 73, 1718–1729 (2010)