

Analysis and Re-use of Videos with Automatic Scene Detection

Lorenzo Baraldi, Costantino Grana, Rita Cucchiara
Dipartimento di Ingegneria “Enzo Ferrari”
Università degli Studi di Modena e Reggio Emilia
Via P. Vivarelli 10, 41125 Modena, Italy
Email: name.surname@unimore.it

Abstract—The advent of modern approaches to education made video the basic media for educating and transmitting knowledge. However, IT tools are still not adequate to allow video content re-use, tagging, annotation and personalization. In this paper we analyze the problem of identifying coherent sequences, called scenes, in order to provide the users with a more manageable editing unit. A simple spectral clustering technique is proposed and compared with state-of-the-art results. We also discuss correct ways to evaluate the performance of automatic scene detection algorithms.

I moderni approcci educativi hanno reso il video il media di riferimento per l'educazione e la trasmissione di conoscenza. Nonostante questo, gli strumenti IT non sono ancora adeguati per permettere il riuso, il tagging, l'annotazione e la personalizzazione dei video in modo efficace. In questo contributo analizziamo il problema di identificare sequenze temporalmente coerenti all'interno di un video (scene), per fornire all'utente finale un video suddiviso in unità più maneggevoli. Viene poi proposta una semplice tecnica di *scene detection* basata sullo *spectral clustering*, che viene poi confrontata con lo stato dell'arte. Si discute infine come valutare correttamente la performance degli algoritmi di *scene detection*.

I. INTRODUCTION

In recent years, the research efforts in video access and video re-use have expanded their interest boundaries beyond traditional fields like news, web entertainment, and sport broadcasting to explore new areas, given the pervasive availability of huge amounts of digital footage. One of such emerging field is surely education that is a key-topic of many international research programs, like the European programs on Smart Communities and the 2020 European Digital Agenda, and that can benefit considerably in accessing the available digital material.

Indeed, many modern approaches to education try to engage the students with technological novelties, such as touch screens [1], hand and body pose recognition [2] or multimedia contents. In particular, Massive Open Online Courses (MOOC) already make use of video as the basic media for educating and transmitting knowledge. Moreover, recent educational projects rethink the concepts of the classical transmission model of the education, towards a socio-cultural-constructivist model where the massive use of video and multimedia content becomes the principal actor in the process of construction of new knowledge centered on the student in strict collaboration between broadcasting bodies, content owners, teachers and the

whole society [3]. For this aim, new instruments should be provided to each level of school for accessing and re-using media contents in different topics, allowing a personalized creation of knowledge, a sharing of multi-cultural practices and the assessment of new social experiences.

In the research project we are involved in, we are developing new solutions for the re-use of educational video production. The goal is to provide efficient tools for students to access the video content, creating their personalized educational experience on specific topics (e.g. geography or art) and across-topics, to share experiences by enriching the footage with user-generated content and data coming from web and social media. In this scenario, even if a huge amount of video from national broadcasting agencies is available and pedagogy researchers are trying to leverage this new possibility in education, the IT tools are still not adequate to allow video content re-use, tagging, annotation and personalization.

Nowadays, people can access video through web or specific apps, but it is difficult to find which section is really the one they want (e.g. a two minute scene withing a two hour program). Even if we know what is the part of interest, extracting and integrating it in our own presentation and re-using it in a suitable manner is still challenging. One basic necessary tool should allow an “access by scene” that improves the level of abstraction from single frame or shot to the scene, i.e. a conceptually meaningful and homogeneous element, composed by more than one shot. Unfortunately, most of the reusable content, owned by broadcast agencies, has not pre-defined sub-units and is not annotated. Therefore, we need accurate scene detection to identify coherent sequences (i.e. scenes) in videos, without asking manual segmentation to editors or publishers. The problem has been approached in the past in the literature with some promising, but not conclusive, results.

We present a novel proposal for scene segmentation, based on spectral clustering, which shows competitive results when compared to state-of-the-art methods. As well, also the broad concept of accuracy should be better defined for scene detection, especially when the goal is not only an algorithm comparison but a concrete result, which should be useful in many applications where a successive human interaction is expected, e.g. for browsing, tagging, selecting etc. In this case, for instance, the precise position of the cut is not important

while skipping a scene and integrating it in another longer, preventing people (in our case students) to find a useful part of the video without seeing all the material, is more important. Thus we compare classical precision/recall measures with a better suited definition of coverage/overflow, which solves frequently observed cases in which the numeric interpretation would be quite different from the expected results by users.

The rest of this paper is organized as follows: Section 2 presents a summary of the existing approaches to scene detection and temporal clustering. In Section 3 we describe our algorithm; in Section 4 we discuss performance evaluation and in Section 5 experimentally evaluate them and show a sample use case.

II. RELATED WORK

Video decomposition techniques aim at partitioning a video into sequences, like shots or scenes, according to semantic or structural criteria. Shots are elementary structural segments that are defined as sequences of frames taken without interruption by a single camera, and shot detection is nowadays regarded as an essentially solved task [5].

Scenes, on the contrary, are often defined as series of temporally contiguous shots characterized by overlapping links that connect shots with similar content [6]. Therefore, the fundamental goal of scene detection algorithms is to identify semantically coherent shots that are temporally close to each other. Most of the existing works can be roughly categorized into two categories: *rule-based methods* and *graph-based methods*. They can rely on visual, audio, and textual features.

Rule-based approaches consider the way a scene is structured in professional movie production. Liu *et al.* [7], for example, propose a visual based probabilistic framework that imitates the authoring process and detects scenes by incorporating contextual dynamics and learning a scene model. In [8], shots are firstly clustered into symbolic groups using spectral clustering. Then, scene boundaries are detected by comparing successive non-overlapping windows of shot labels using a modified version of the Needleman-Wunsh algorithm, that considers the visual similarity of shot clusters and the frequency of sequential labels in the video. Of course, the drawback of this kind of methods is that they tend to fail when directors intentionally break film-editing rules, or when two adjacent scenes are similar and follow the same rules.

In graph-based methods, instead, shots are arranged in a graph representation and then clustered by partitioning the graph. The Shot Transition Graph (STG), proposed in [9], is one of the most used models in this category: here each node represents a shot and the edges between the shots are weighted by shot similarity. In [10], color and motion features are used to represent shot similarity, and the STG is then split into subgraphs by applying the normalized cuts for graph partitioning.

More recently, Sidiropoulos *et al.* [11] extended the Shot Transition Graph using multimodal low-level and high-level features. To this aim, multiple STGs are constructed, one for each kind of feature, and then a probabilistic merging process

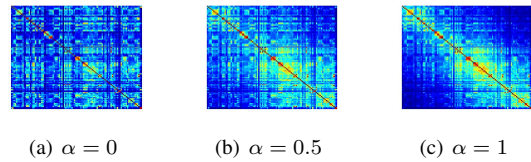


Fig. 1. Effect of α on similarity matrix κ_{ij} . Higher values of α enforce connections between near shots and increase the quality of the detected scenes (best viewed in color).

is used to combine their results. The features used include visual features, such as HSV histograms, outputs of visual concept detectors trained using the Bag of Words approach, and audio features, like background conditions classification results, speaker histogram, and model vectors constructed from the responses of a number of audio event detectors.

III. A SPECTRAL CLUSTERING APPROACH

Our scene detection method generates scenes by grouping adjacent shots. Shots are described by means of color histograms, hence relying on visual features only: given a video, we compute a three-dimensional histogram of each frame, by quantizing each RGB channel in eight bins, for a total of 512 bins. Then, we sum histograms from frames belonging to the same shot, thus obtaining a single L_1 -normalized histogram for each shot.

In contrast to other approaches that used spectral clustering for scene detection, we build a similarity matrix that jointly describes appearance similarity and temporal proximity. Its generic element κ_{ij} , defines the similarity between shots \mathbf{x}_i and \mathbf{x}_j as

$$\kappa_{ij} = \exp\left(-\frac{d_1^2(\psi(\mathbf{x}_i), \psi(\mathbf{x}_j)) + \alpha \cdot d_2^2(\mathbf{x}_i, \mathbf{x}_j)}{2\sigma^2}\right) \quad (1)$$

where $\psi(\mathbf{x}_i)$ is the normalized histogram of shot \mathbf{x}_i , d_1^2 is the Bhattacharyya distance and $d_2^2(\mathbf{x}_i, \mathbf{x}_j)$ is the normalized temporal distance between shot \mathbf{x}_i and shot \mathbf{x}_j , while the parameter α tunes the relative importance of color similarity and temporal distance. To describe temporal distance between frames, $d_2^2(\mathbf{x}_i, \mathbf{x}_j)$ is defined as

$$d_2^2(\mathbf{x}_i, \mathbf{x}_j) = \frac{|m_i - m_j|}{l} \quad (2)$$

where m_i is the index of the central frame of shot \mathbf{x}_i , and l is the total number of frames in the video. The spectral clustering algorithm is then applied to the similarity matrix, using the Normalized Laplacian and the maximum eigen-gap criterion to select k :

$$k = \arg \max (\lambda_{i+1} - \lambda_i) + 1 \quad (3)$$

where λ_i is the i -th eigenvalue of the Normalized Laplacian. As shown in Fig. 1, the effect of applying increasing values of α to the similarity matrix is to raise the similarities of adjacent shots, therefore boosting the temporal consistency of the resulting groups. Of course, this does not guarantee



Fig. 2. Two consecutive scenes from the RAI dataset.

a completely temporal consistent clustering (i.e. some clusters may still contain non-adjacent shots); at the same time, too high values of α would lead to a segmentation that ignores color dissimilarity. The final scene boundaries are created between adjacent shots that do not belong to the same cluster.

IV. EVALUATING SCENE SEGMENTATION

The first possibility to evaluate the results of a scene detection algorithm is to count correctly and wrongly detected boundaries, without considering the temporal distance between a ground truth cut and the nearest detected cut. The most used measures in this context are precision and recall, together with the F-Score measure, that summarizes both. Precision is the ratio of the number of correctly identified scenes boundaries to the total number of scenes detected by the algorithm. Recall is the ratio of the number of correctly identified boundaries to the total number of scenes in the ground truth.

Of course this kind of evaluation does not discern the seriousness of an error: if a boundary is detected one shot before or after its ground truth position, an error is counted in recall as if the boundary was not detected at all, and in precision as if the boundary was put far away. This issue appears to be felt also by other authors, with the result that sometimes a tolerance factor is used. For example, [10] uses a *best match* method with a sliding window of 30 seconds, so that a detected boundary is considered correct if it matches a ground truth boundary in the sliding window.

To deal with these problems, Vendrig *et al.* [12] proposed the Coverage and Overflow measures. Coverage \mathcal{C} measures the quantity of shots belonging to the same scene correctly grouped together, while Overflow \mathcal{O} evaluates to what extent shots not belonging to the same scene are erroneously grouped together. Formally, given the set of automatically detected scenes $\mathbf{s} = [s_1, s_2, \dots, s_m]$, and the ground truth $\tilde{\mathbf{s}} = [\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_n]$, where each element of \mathbf{s} and $\tilde{\mathbf{s}}$ is a set

TABLE I
PERFORMANCE COMPARISON ON THE RAI DATASET USING THE
BOUNDARY LEVEL MEASURES (F-SCORE)

| Video | Our method | Method in [8] | Method in [11] |
|----------|------------|---------------|----------------|
| V_1 | 0.12 | 0.25 | 0.29 |
| V_2 | 0.36 | 0.00 | 0.30 |
| V_3 | 0.37 | 0.13 | 0.31 |
| V_4 | 0.30 | 0.10 | 0.22 |
| V_5 | 0.44 | 0.00 | 0.36 |
| V_6 | 0.18 | 0.00 | 0.36 |
| V_7 | 0.18 | 0.00 | 0.13 |
| V_8 | 0.10 | 0.13 | 0.21 |
| V_9 | 0.25 | 0.00 | 0.21 |
| V_{10} | 0.23 | 0.26 | 0.19 |
| Average | 0.25 | 0.09 | 0.26 |

of shot indexes, the coverage \mathcal{C}_t of scene \tilde{s}_t is proportional to the longest overlap between \mathbf{s}_i and \tilde{s}_t :

$$\mathcal{C}_t = \frac{\max_{i=1, \dots, m} \#(\mathbf{s}_i \cap \tilde{s}_t)}{\#(\tilde{s}_t)} \quad (4)$$

where $\#(\mathbf{s}_i)$ is the number of shots in scene \mathbf{s}_i . The overflow of a scene \tilde{s}_t , \mathcal{O}_t , is the amount of overlap of every \mathbf{s}_i corresponding to \tilde{s}_t with the two surrounding scenes \tilde{s}_{t-1} and \tilde{s}_{t+1} :

$$\mathcal{O}_t = \frac{\sum_{i=1}^m \#(\mathbf{s}_i \setminus \tilde{s}_t) \cdot \min(1, \#(\mathbf{s}_i \cap \tilde{s}_t))}{\#(\tilde{s}_{t-1}) + \#(\tilde{s}_{t+1})} \quad (5)$$

The computed per-scene measures can then be aggregated into values for an entire video as follows:

$$\mathcal{C} = \sum_{t=1}^n \mathcal{C}_t \cdot \frac{\#(\tilde{s}_t)}{\sum \#(\tilde{s}_i)}, \quad \mathcal{O} = \sum_{t=1}^n \mathcal{O}_t \cdot \frac{\#(\tilde{s}_t)}{\sum \#(\tilde{s}_i)}. \quad (6)$$

Finally, an F-Score measure can be defined to combine Coverage and Overflow in a single measure, by taking the harmonic mean of \mathcal{C} and $1 - \mathcal{O}$.

V. EXPERIMENTAL EVALUATION

We evaluate the aforesaid measures and algorithms on a collection of ten challenging broadcasting videos from the Rai Scuola video archive¹, mainly documentaries and talk shows. Shots have been obtained running the state of the art shot detector of [13] and manually grouped into scenes by a set of human experts to define the ground truth. Our dataset and the corresponding annotations are available for download at <http://imagelab.ing.unimore.it/files/RaiSceneDetection.zip>.

We reimplemented the approach in [8] and used the executable of [11] provided by the authors². The threshold Th of [8] was selected to maximize the performance on our dataset.

Tables I and II compare the two different approaches using Boundary level and Shot level performance measures. As show in Table I, detected boundaries rarely correspond to ground truth boundaries exactly, therefore leading to poor results in terms of precision and recall, even when considering a recent and state-of-the-art approach like [11].

¹<http://www.scuola.rai.it>

²<http://mklab.iti.gr/project/video-shot-segm>

TABLE II
PERFORMANCE COMPARISON ON THE RAI DATASET USING THE SHOT LEVEL MEASURES (COVERAGE, OVERFLOW AND F-SCORE)

| Video | Spectral Clustering | | | Chasanis <i>et al.</i> [8] | | | Sidiropoulos <i>et al.</i> [11] | | |
|-----------------|---------------------|------|------|----------------------------|------|------|---------------------------------|------|------|
| | F-Score | C | O | F-Score | C | O | F-Score | C | O |
| V ₁ | 0.64 | 0.81 | 0.48 | 0.70 | 0.64 | 0.24 | 0.72 | 0.84 | 0.37 |
| V ₂ | 0.68 | 0.61 | 0.22 | 0.36 | 0.80 | 0.77 | 0.59 | 0.85 | 0.55 |
| V ₃ | 0.65 | 0.68 | 0.38 | 0.58 | 0.73 | 0.52 | 0.58 | 0.90 | 0.57 |
| V ₄ | 0.74 | 0.69 | 0.22 | 0.50 | 0.65 | 0.60 | 0.33 | 0.94 | 0.80 |
| V ₅ | 0.77 | 0.68 | 0.11 | 0.25 | 0.93 | 0.86 | 0.66 | 0.76 | 0.41 |
| V ₆ | 0.51 | 0.37 | 0.17 | 0.18 | 0.89 | 0.90 | 0.71 | 0.77 | 0.34 |
| V ₇ | 0.30 | 0.97 | 0.82 | 0.37 | 0.70 | 0.75 | 0.51 | 0.78 | 0.62 |
| V ₈ | 0.59 | 0.53 | 0.33 | 0.62 | 0.57 | 0.32 | 0.45 | 0.88 | 0.70 |
| V ₉ | 0.67 | 0.55 | 0.15 | 0.27 | 0.87 | 0.84 | 0.43 | 0.92 | 0.72 |
| V ₁₀ | 0.57 | 0.42 | 0.12 | 0.54 | 0.91 | 0.62 | 0.44 | 0.94 | 0.71 |
| Average | 0.61 | 0.63 | 0.30 | 0.44 | 0.77 | 0.64 | 0.54 | 0.86 | 0.58 |

As expected, the two measures behave differently and there is not a complete agreement among them: [8] performs worse than the other two methods according to both measures, while [11] performs equal to the spectral clustering approach with boundary level measures, but slightly worse than the spectral clustering approach according to shot level measures.

Detected scenes, finally, can be used as an input for video browsing or re-using software. As an example, we built a web-based browsing interface for broadcasting videos (see Figure 3) where users can visualize a summary of the content by means of the extracted scenes. Scenes are represented with key-frames in a time-line fashion. To ease the browsing even more, tags obtained with automatic concept annotation are reported under each shot. Users can jump from one part of the video to another by clicking on the corresponding scene or shot. A demo of our interface is available at <http://imabelab.ing.unimore.it/scenedemo>.

VI. CONCLUSIONS

We investigated the problem of evaluating scene detection algorithms with tests conducted on two different performance measures and on three different and recent approaches to scene segmentation. Results show that the problem of scene detection is still far from being solved, and that simple approaches like the suggested spectral clustering technique can sometimes achieve equivalent or better results than more complex methods.

ACKNOWLEDGMENTS

This work was carried out within the project “Città educante” (ctn01_00034_393801) of the National Technological Cluster on Smart Communities cofunded by the Italian Ministry of Education, University and Research - MIUR.

REFERENCES

- [1] J. K. Battenberg and J. B. Merbler, “Touch screen versus keyboard: A comparison of task performance of young children.” *Journal of Special Education Technology*, vol. 10, no. 2, pp. 24–28, 1989.
- [2] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara, “Gesture recognition in ego-centric videos using dense trajectories and hand segmentation,” in *Proc. of 10th IEEE Embedded Vision Workshop (EVW)*, Columbus, Ohio, Jun. 2014.
- [3] M. F. Mascolo, “Beyond student-centered and teacher-centered pedagogy: Teaching and learning as guided participation,” *Pedagogy and the Human Sciences*, vol. 1, no. 1, pp. 3–27, 2009.

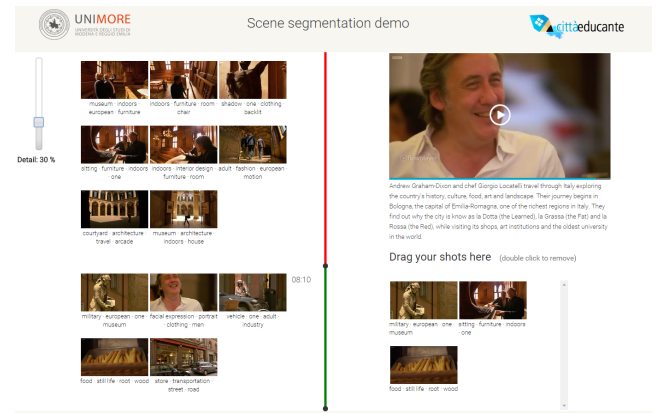


Fig. 3. Effective video browsing using our algorithm. Users can visualize a summary of the content by means of the extracted scenes.

- [4] B. T. Truong and S. Venkatesh, “Video abstraction: A systematic review and classification,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 3, no. 1, p. 3, 2007.
- [5] A. Hanjalic, R. L. Lagendijk, and J. Biemond, “Automated high-level movie segmentation for advanced video-retrieval systems,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 580–588, 1999.
- [6] C. Liu, D. Wang, J. Zhu, and B. Zhang, “Learning a Contextual Multi-Thread Model for Movie/TV Scene Segmentation,” *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 884–897, 2013.
- [7] V. T. Chasanis, C. Likas, and N. P. Galatsanos, “Scene detection in videos using shot clustering and sequence alignment,” *IEEE Transactions on Multimedia*, vol. 11, no. 1, pp. 89–100, 2009.
- [8] M. M. Yeung, B.-L. Yeo, W. H. Wolf, and B. Liu, “Video browsing using clustering and scene transitions on compressed sequences,” in *IS&T/SPIE’s Symposium on Electronic Imaging: Science & Technology*, 1995, pp. 399–413.
- [9] Z. Rasheed and M. Shah, “Detection and representation of scenes in videos,” *IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1097–1105, 2005.
- [10] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso, “Temporal video segmentation to scenes using high-level audiovisual features,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 8, pp. 1163–1177, 2011.
- [11] J. Vendrig and M. Worring, “Systematic evaluation of logical story unit segmentation,” *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 492–499, 2002.
- [12] E. Apostolidis and V. Mezaris, “Fast Shot Segmentation Combining Global and Local Visual Descriptors,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 6583–6587.