SEGMENTATION OF TV SHOWS INTO SCENES USING SPEAKER DIARIZATION AND SPEECH RECOGNITION

Hervé Bredin

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay Cedex, France
bredin@limsi.fr

ABSTRACT

We investigate the use of speaker diarization (SD) and automatic speech recognition (ASR) for the segmentation of audiovisual documents into scenes. We introduce multiple monomodal and multimodal approaches based on a state-of-the-art algorithm called *generalized scene transition graph* (GSTG). First, we extend the latter with the use of semantic information derived from both SD and ASR. Then, multimodal fusion of color histograms, SD and ASR is investigated at various point of the GSTG pipeline (early, late or intermediate fusion). Experiments driven on a few episodes of a popular TV show indicate that SD and ASR can be successfully combined with visual information and bring an additional $+11\%$ relative increase in terms of F_1-measure for scene boundary detection over the state-of-the-art baseline.

Index Terms— scene boundary detection, speaker diarization, speech recognition, scene transition graph, multimodal fusion

1. INTRODUCTION

From content-based multimedia indexing to automatic video summarization, most existing applications dealing with multimedia analysis rely on a preliminary step dedicated to temporal segmentation. It is typically achieved in a hierarchical manner: consecutive video frames are grouped into camera shots (a.k.a. shot boundary detection), then combined into higher-level *semantic* segments such as stories in TV broadcast news. Semantic segmentation of videos allows for easier and faster browsing in ever-growing collections. For instance, it can be used to add chapter markers in long videos or as the basis for automatic video summarization.

Segmentation of video into *scenes* gained a lot of attention in the last decade. Methods proposed in the literature are often domain-specific. In [1], the authors use explicit rules coming from the audiovisual production domain to achieve segmentation into scenes. Other approaches rely on strong a priori knowledge on the segmented videos and will only perform well on specific type of content (such as sports events [2] for instance). Most existing methods do not perform well on heterogeneous corpora. In particular, scene boundary detection can be tricky for movies or TV shows since their construction obeys to subjective (artistic) criterions.

This paper offers a complete redesign of our previous work focusing on the segmentation of TV shows into scenes [3]. Defining the concept of a *scene* is a difficult problem in itself and we could find nearly as many definitions as there are researchers working on this very subject. Some consider that scenes have nothing to do with semantics [4] while others assert the contrary [5]. We choose to consider that a scene is composed of a set of consecutive shots with the following properties:

Temporal continuity: a scene should describe an event in a continuous manner. A new scene should be detected for every flashback or flashforward, for instance. Similarly, a change in setting is usually a strong indication that time flew by between two scenes. This is a fully *objective* property.

Semantic coherence: this one is *subjective*. Even in case of temporal continuity, it might happen that a completely new story is beginning (with the introduction of a new character or an external event for instance) at some time. When this happens, a new scene should be detected.

Most TV shows narrate the story of a relatively small number of recurring characters. For instance, the show called *Ally McBeal* describes the every day life of a female lawyer in her thirties and her relationships with her colleagues and friends at the law firm *Cage & Fish*. Among other things, dialogues between characters are a mean to describe and make the story evolve. Multiple stories are narrated in parallel, describing various facets of their lives.

This is why we investigate the use of speaker diarization (SD) and automatic speech recognition (ASR) in addition to visual cues to achieve better segmentation results. Using both visual and audio low-level features for scene boundary detection is not a new idea [4, 6], but as far as we know, successfully combining SD and ASR with visual information is.

Section 2 offers a quick overview of a state-of-the-art algorithm for scene boundary detection. Our first contribution is described in Section 3 where we show how speaker diarization and speech recognition can be used in this framework. Three multimodal fusion approaches are introduced in Section 4: they differ in how early audio and visual sources of information are combined, in the scene boundary detection process. The experimental framework is summarized in Section 5, in which both monomodal and multimodal approaches are evaluated. Finally, Section 6 concludes the paper.

2. SCENE BOUNDARY DETECTION

In this section, we describe the so-called *scene transition graph* (STG) approach for scene boundary detection introduced in 1998 by *Yeung et al.* [7] and its recent extension by *Sidiropoulos et al.* called generalized STG [6]. They both assume that shot boundaries are readily available and consider scene boundary detection as the following classification problem: *"for each shot boundary, is it also a boundary between two scenes?"* Our own monomodal and multimodal approaches for scene boundary detection are based on generalized STG and this assumption.

2.1. Scene transition graph

Let us denote d_{ij} the dissimilarity of shots i and j and t_{ij} their temporal distance. As will be described in Section 3, the dissimilarity of two shots can be as simple (and low-level) as the distance between color histograms extracted from the two shots; or may drive more semantic information related to characters or dialogues. Assuming $i < j$, the temporal distance t_{ij} is simply defined as the duration between the end of shot i and the beginning of shot j. Dissimilarity and temporal distance are then combined:

$$D_{ij} = \begin{cases} d_{ij} & \text{if } t_{ij} < \Delta_t \\ +\infty & \text{otherwise} \end{cases} \quad (1)$$

and a complete-link agglomerative clustering approach allows to group shots together based on this new distance D. The agglomerative clustering ends when the distance between the two closest clusters is higher than a threshold Δ_d. Overall, this process allows to group semantically similar shots together, as long as they are not too far away from each other in the video. In the example shown in Figure 1, 11 consecutive shots are clustered into 5 groups using this approach.

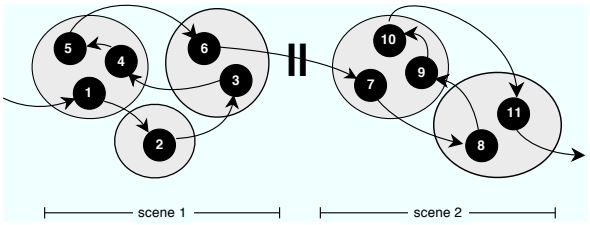

Fig. 1. Scene transition graph.

Consequently, the so-called scene transition graph (STG) is built with one vertex per shot and one edge connecting each pair of consecutive shots. Cut-edges (that is edges whose removal splits the whole graph into 2 disjoint connected components) are then detected and the corresponding shot boundary is marked as a scene boundary. In Figure 1, the edge between vertices #6 and #7 is found to be a cut-edge. Therefore, the boundary between shots #6 and #7 is marked as a scene boundary.

2.2. Generalized STG

Each pair of values (Δ_d, Δ_t) leads to a different set of detected scene boundaries. The optimal values (i.e. leading to the best set of scene boundaries) are dependent on the video. One smart way of (partially) removing the need for fine-tuning these parameters was proposed in [6] with the introduction of the Generalized Scene Transition Graph (GSTG). The idea is to generate a large set of STGs by selecting random values for Δ_d and Δ_t, and keep track of the percentage of STGs that found every shot boundary to be a scene boundary. In Figure 2, shot boundaries with percentage p higher than a threshold θ are marked as scene boundaries (vertical lines).

Sidiropoulos et al. found this approach to give better performance and our own preliminary experiments confirmed this observation. Not only is it easier to fine-tune one single parameter (the threshold θ) instead of two of them (Δ_d and Δ_t), but the GSTG approach also gives the best (biased) optimal performance when parameters are tuned directly on the test set.

Rather than randomly selecting values for Δ_d and Δ_t, our own implementation generates an exhaustive set of STGs using all possible pairs of values for Δ_d and Δ_t (in a predefined 2-dimensional grid). This has the advantage over the original approach to be deterministic and as such leads to reproducible results.

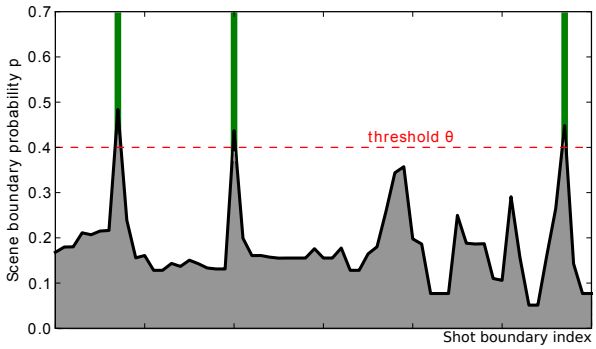

Fig. 2. Scene boundary probability

3. MONO-MODAL APPROACHES

In this section, we describe two novel monomodal approaches for segmentation of videos into scenes. Both are based on the GSTG algorithm proposed by *Sidiropoulos et al.* but differ on the computation of the dissimilarity between shots.

For comparison purposes and later multimodal approaches (in Section 4), a baseline monomodal system based on HSV color histograms was also implemented. Color histograms (10x10x10 bins) are extracted every second and the dissimilarity d_{ij}^{HSV} between two shots i and j is defined as the minimum Manhattan distance between all possible pairs of histograms from these two shots.

However, one cannot expect to solve this problem using low-level descriptors only. Therefore, as summarized in Figure 3, aside from the baseline HSV color histograms, we propose to use higher-level semantic descriptors extracted from the soundtrack.

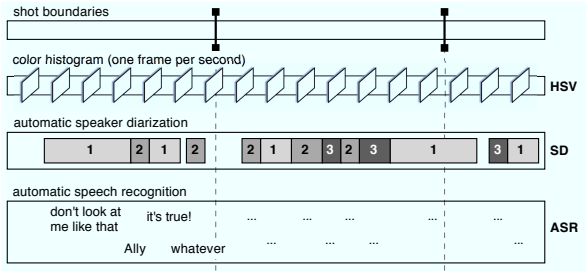

Fig. 3. Set of available modalities

3.1. Speaker diarization

Speaker diarization is the process of partitioning the audio stream into homogeneous segments, based on the identity of the speaker. **SD** timeline in Figure 3 shows an example of the output of such a system: speech turns are detected and then labelled with a unique speaker identifier (1, 2 or 3). Zero, one or more speakers may speak during each shot.

In order to compute a unique dissimilarity d_{ij} between each pair of shots (i, j), we propose to use the TF-IDF paradigm, borrowed from the text document retrieval community. Each shot s is described by a D_{SD}-dimensional feature vector $X(s)$ where D_{SD} is the total number of speakers in the video and $\text{X}_d(s) = \text{TF}_d(s) \times \text{IDF}_d$ for $d \in \{1 \dots D_{\text{SD}}\}$:

- Inverse document frequency (IDF) is defined by $\text{IDF}_d = \log(N/N_d)$ where N is the number of shots in the video and N_d the number of shots during which speaker d actually speaks.
- Term-frequency (TF) is defined by $\text{TF}_s^d = L_d(s)/L(s)$ where $L(s)$ is the duration of shot s and $L_d(s)$ is the speech duration of speaker d in shot s.

The SD-based dissimilarity d_{ij}^{SD} between shots i and j is defined as the cosine distance between their respective TF-IDF feature vectors.

3.2. Speech recognition

In order to bring even more semantic information to the game, we also propose to use the output of an automatic speech recognition (ASR) system as another complementary modality. The ASR output is processed by TreeTagger [8] in order to extract the lemma of each recognized word. Each shot s is then described by another D_{ASR}-dimensional TF-IDF feature vector, where D_{ASR} is the total number of unique lemmas recognized by the ASR system:

- Inverse document frequency (IDF) is defined by $\text{IDF}_d = \log(N/M_d)$ where N is the number of shots in the video and M_d the number of shots containing at least one occurrence of dth lemma.
- Term-frequency (TF) is defined by $\text{TF}_s^d = W_d(s)/W(s)$ where $W_d(s)$ is the number of occurrences of dth lemma in shot s and $W(s)$ is the number of words recognized in shot s.

The ASR-based dissimilarity d_{ij}^{ASR} between shots i and j is defined as the cosine distance between their respective TF-IDF feature vectors.

4. MULTIMODAL FUSION

In this section, we present several ways of combining the above monomodal approaches into a (hopefully better) multimodal scene boundary detection system. In particular, we distinguish three approaches, differing in how early in the GSTG-based detection pipeline the multimodal combination happens. Figure 4 summarizes it all.

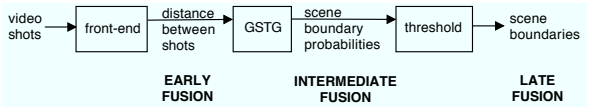

Fig. 4. Early vs. intermediate vs. late fusion

4.1. Early fusion

Early fusion is performed once dissimilarities between shots d_{ij} are available from each combined modality. It consists in their linear combination and can be summarized as follows:

$$d_{ij} = w_{\text{HSV}} \cdot d_{ij}^{\text{HSV}} + w_{\text{SD}} \cdot d_{ij}^{\text{SD}} + w_{\text{ASR}} \cdot d_{ij}^{\text{ASR}} \quad (2)$$

with $w_{\text{HSV}} + w_{\text{SD}} + w_{\text{ASR}} = 1$. As described in Section 5, a leave-one-out cross-validation paradigm is used to estimate the best weight for each modality.

Obviously, monomodal dissimilarities are first normalized so that they all share a similar range of values and none of them outweighs the others. We investigated multiple normalization techniques (min/max, z-score, TanH) but only report on the one that proved to be the best: gaussian normalization. It consists in modifying the original distribution of d_{ij} so that it is as close as possible to a normal distribution.

4.2. Intermediate fusion

Intermediate fusion consists in a linear combination of the scene boundary probabilities p generated by the monomodal GSTGs. For each shot boundary, its multimodal scene boundary probability is defined as follows:

$$p = w_{\text{HSV}} \cdot p_{\text{HSV}} + w_{\text{SD}} \cdot p_{\text{SD}} + w_{\text{ASR}} \cdot p_{\text{ASR}} \quad (3)$$

with the same constraint on weights ($w_{\text{HSV}} + w_{\text{SD}} + w_{\text{ASR}} = 1$). This particular approach was the one proposed by the authors of [6] – even though they combine modalities that are different from ours.

4.3. Late fusion

We define two simple late fusion approaches: intersection $\cap$ and union $\cup$. In intersection $\cap$ fusion, a shot boundary is marked as a scene boundary if all combined monomodal approaches marked it as one. In union $\cup$ fusion, it is marked as a scene boundary if at least one monomodal approach marked it as one. These are very simple ways of performing late fusion and will serve as baselines for the other fusion approaches.

5. EXPERIMENTS

In order to perform experiments, we acquired the first season of the *Ally McBeal* TV series. We manually annotated the first eight episodes with shot and scene boundaries – for a total duration of around 5 hours of videos, 5564 shots and 306 scenes. HSV color histograms were extracted every second using the OpenCV library. Speaker diarization and speech recognition were automatically generated using the LIMSI speaker diarization and automatic speech recognition tools [9].

We consider the segmentation problem as a boundary detection problem and therefore rely on precision, recall and their combination into F_1-measure. A detected boundary is correct if it has the exact same position as a groundtruth boundary (and incorrect otherwise) – no temporal tolerance is allowed.

Since only eight episodes are annotated, the evaluation protocol follows the *leave-one-out* cross-validation paradigm. Optimal parameters (i.e. maximizing the F_1-measure) are obtained by tuning the segmentation algorithms using seven episodes (training set) and are applied on the remaining episode (validation set) – this process being repeated for each episode. The reported value is computed as the average of values obtained from the eight combinations.

5.1. Mono-modal experiments

Table 1 summarizes the performance of monomodal approaches. The baseline system – based on HSV color histograms only – obtains by far the best results with an F_1-measure of 0.487.

	Precision	Recall	F_1-measure	# Scenes
HSV	**0.447**	0.566	**0.487**	403
SD	0.157	0.562	0.240	1136
ASR	0.105	0.572	0.175	1751

Table 1. Performance of monomodal approaches

Both SD- and ASR-based approaches tend to detect far too many scene boundaries (resp. 1100+ and 1700+ detected scenes when the actual correct number is close to 300). This behavior leads to very low precision values (resp. 0.16 and 0.10).

This can be explained by the way SD and ASR distances are computed and is especially true for the latter. Indeed, since two shots rarely have more than one or two words in common, their ASR distance tends to be very close to 1. Therefore, the agglomerative clustering process usually stops very early (depending on threshold Δ_d), and generates a long scene transition graph with lots of clusters made of only one shot (and lots of cut-edges).

5.2. Multi-modal experiments

Though SD- and ASR-based approaches have worse performance than the HSV baseline, multimodal fusion results reported in Table 2 show they can lead to much more accurate scene detection when combined with it.

Fusion	Precision	Recall	F_1-measure
HSV (baseline)	0.447	0.566	**0.487**
HSV ∩ SD	**0.598**	0.357	0.438
HSV ∩ SD ∩ ASR	**0.606**	0.242	0.341
HSV ∪ SD	0.180	**0.770**	0.288
HSV ∪ SD ∪ ASR	0.121	**0.851**	0.210
d(HSV) + d(SD)	0.445	0.599	0.499
d(HSV) + d(SD) + d(ASR)	0.445	0.599	0.499
p(HSV) + p(SD)	0.484	0.555	**0.510**
p(HSV) + p(SD) + p(ASR)	0.488	0.622	**0.539**

Table 2. Performance of fusion approaches

As expected, intersection $\cap$ (resp. union $\cup$) fusion greatly improves precision (resp. recall), yet at the expense of recall (resp. precision). If increasing the F_1-measure is the objective, $\cap$ and $\cup$ late fusion approaches must be avoided.

As far as early (distance) fusion is concerned, Table 2 shows that one can significantly improve the HSV baseline by combining it with the one based on speaker diarization (F_1-measure increases from 0.487 to 0.499). On average, $w_{\text{HSV}} \approx 0.9$ and $w_{\text{SD}} \approx 0.1$ are the optimal weights obtained by cross-validation. Unfortunately, adding ASR does not help as it is always given a null weight.

The fusion paradigm that gives the best result is the intermediate one. Combining HSV with SD greatly improves precision while keeping recall at a similar level – F_1-measure therefore increases from 0.487 to 0.510, which is already better than any early or late fusion approaches we tested. Then, adding ASR to the game significantly improves recall (from 0.555 to 0.622). All in all, intermediate fusion of HSV, SD and ASR gives an 11% improvement in terms of F_1-measure, up to 0.539.

6. CONCLUSIONS

We investigated the use of speaker diarization (SD) and automatic speech recognition (ASR) for the segmentation of audiovisual documents into scenes. We introduced multiple monomodal and multimodal approaches based on *generalized scene transition graphs*. First, we extended the latter with the use of semantic information derived from both SD and ASR. We found that they both tend to detect far too many scene boundaries. However, we then showed that they can efficiently be combined with visual information to improve upon the state-of-the-art baseline. In particular, the intermediate fusion approach gives the better result with a relative $+11\%$ increase of F_1-measure ($+9\%$ for precision and $+9\%$ for recall on average).

We made our best to implement the state-of-the-art GSTG approach as close to its description in [6] as possible. However, since no standard evaluation dataset has emerged yet in the community, it is very difficult to compare with other works. Therefore, we share our own manual annotations for other to use[1].

As far as future work is concerned, it should obviously be possible to use additional modalities that drive even more semantic information [6] and combine them altogether at intermediate level. However, we think that we might have reached the limits of the GSTG approach and plan on investigating other graph-based approaches. Furthermore, segmentation into scenes is not an end in itself and we expect to use the results presented in this paper in a larger framework dedicated to automatic video summarization and semantically-driven video browsing.

7. ACKNOWLEDGMENT

This work was partly realized as part of the Quaero Program and the QCompere project, respectively funded by OSEO (French State agency for innovation) and ANR (French national research agency).

8. REFERENCES

[1] Wallapak Tavanapong and Junyu Zhou, "Shot Clustering Techniques for Story Browsing," *IEEE Transactions on Multimedia*, vol. 6, no. 4, pp. 517–527, August 2004.

[2] Lexing Xie, Peng Xu, Shih-Fu Chang, Ajay Divakaran, and Huifang Sun, "Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models," *Pattern Recognition Letters - Video computing*, vol. 25, pp. 767–775, May 2004.

[3] Philippe Ercolessi, Hervé Bredin, Christine Sénac, and Philippe Joly, "Segmenting TV Series into Scenes using Speaker Diarization," in *WIAMIS 2011:, 12th International Workshop on Image Analysis for Multimedia Interactive Services*, Delft, The Netherlands, April 2011.

[4] Hari Sundaram and Shih-Fu Chang, "Computable Scenes and Structures in Films," *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 482 – 491, December 2002.

[5] Songhao Zhu and Yuncai Liu, "Video Scene Segmentation and Semantic Representation Using a Novel Scheme," *Multimedia Tools and Applications*, vol. 42, no. 2, pp. 183–205, April 2009.

[6] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bugalho, and Isabel 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, August 2011.

[7] Minerva Yeung, Boon-Lock Yeo, and Bede Liu, "Segmentation of Video by Clustering and Graph Analysis," *Computer Vision and Image Understanding*, vol. 71, no. 1, pp. 94–109, July 1998.

[8] Helmut Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *Proceedings of the International Conference on New Methods in Language Processing*, 1994, pp. 44–49.

[9] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, vol. 37, no. 1-2, pp. 89–109, 2002.

[1] Available at `http://tinyurl.com/AllyAnnotations`