SEGMENTING TV SERIES INTO SCENES USING SPEAKER DIARIZATION

Philippe Ercolessi, Christine Sénac, Philippe Joly

IRIT
118 Route de Narbonne
Toulouse, France

Hervé Bredin

Spoken Language Processing Group
CNRS-LIMSI, BP 133
Orsay, France

ABSTRACT

In this paper, we propose a novel approach to perform scene segmentation of TV series. Using the output of our existing speaker diarization system, any temporal segment of the video can be described as a binary feature vector. A straightforward segmentation algorithm then allows to group similar contiguous speaker segments into scenes. An additional visual-only color-based segmentation is then used to refine the first segmentation. Experiments are performed on a subset of the *Ally McBeal* TV series and show promising results, obtained with a rule-free and generic method. For comparison purposes, test corpus annotations and description are made available to the community.

1. INTRODUCTION

Because this is a mandatory pre-processing step for most applications dealing with multimedia analysis, temporal video segmentation has been studied extensively.

Traditionally, a hierarchical approach is adopted to perform the analysis of the video structure. First, at the bottom of the structure, consecutive video frames are grouped into camera shots. Then, several works have attempted to find a *semantic* structure at a higher level by grouping together adjacent shots into scenes.

In [1], the authors use more or less explicit rules coming from the audiovisual production domain to achieve scene segmentation. Scene boundary detection is based on a graph-based representation of the video in [2], on statistical learning in [3] or audiovisual features in [4].

Overall, the methods proposed in the state-of-the-art do not perform well on heterogeneous corpora. They use a priori knowledge on the video content or genre and each one has its own definition of a scene: some consider that scenes do not have to be related to semantics [4] while others assert the contrary [5]. Yet, scenes can be detected from specific types of programs with a stable structure such as broadcast news or sports events [3]. On the other hand, this task can be tricky for movies or television (TV) series because it obeys to subjective criterions.

In this paper, we present a novel unsupervised approach for scene boundary detection in TV series.

Among the multiple definitions of a *scene*, we choose to consider that a scene is composed of a set of shots showing a spatio-temporal continuity. Thereby, a scene boundary occurs either when the place changes, or when the time of action changes between two consecutive shots (for instance, when the previous shot shows a character at night, and the current one shows this same character during the day).

Most TV series narrate the story of a relatively small number of recurring characters. Dialogues between characters is a mean to describe and make the story evolve. Moreover, multiple sub-stories are usually narrated in parallel, describing various facets of the main character's lives.

It should therefore be possible to partially split a whole episode into scenes based on the knowledge of who is speaking and when.

Thus, our method is based on the output of our speaker diarization system [6]. Speech segments are grouped into scenes following a principle described in Section 2. As speaker-based segmentation does not always match the actual scene segmentation, we also benefit from a color-based segmentation (Section 3) in order to enhance the scene boundaries (Section 4). Finally, experiments are described in Section 5.

2. SPEAKER-BASED SEGMENTATION

Our approach is divided into two steps: a speaker diarization followed by the segmentation into scenes.

2.1. Speaker diarization and binary representation

Speaker diarization is the process of segmenting an audio stream and clustering resulting segments in different speakers. We use the system described in [6] to obtain a labelled segmentation as shown in Figure 1.

Throughout this process, speech segments emanating from the same speaker are gathered and annotated with the same label. Let D be the number of different speakers found in a document ($D = 3$ in Figure 1).

Consequently, any audio segment can be represented as a D-dimensional binary feature vector $\mathbf{x} \in \{0,1\}^D$, with $\mathbf{x} =$

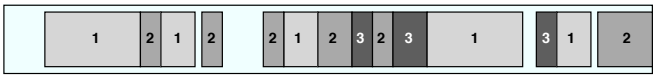

Fig. 1. Speaker diarization – three different speakers (labelled 1, 2 and 3) were detected.

$[x_1, x_2, \ldots, x_D]$ where

$$x_i = \begin{cases} 1 & \text{if speaker } i \text{ speaks during segment} \\ 0 & \text{otherwise} \end{cases}$$

The binary feature vector $\mathbf{x}$ extracted from three audio segments at various temporal positions is illustrated in Figure 2.

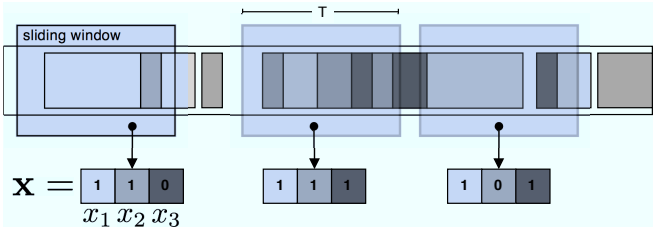

Fig. 2. Binary description

2.2. Segmentation

Let us consider a sliding window of duration T. We denote $\mathbf{x}_t$ the binary feature vector extracted from the window starting at time t. The proposed segmentation relies entirely on this binary description and can be summarized in pseudo-code as in Algorithm 1.

Algorithm 1 Segmentation based on binary feature vectors. $\mathcal{S}$ is the list of scenes and δ the step of the sliding window.

1: $\mathcal{S} \Leftarrow \emptyset$
2: $t_0 \Leftarrow 0$
3: $t \Leftarrow t_0 + \delta$
4: **while** $d(\mathbf{x}_{t_0}, \mathbf{x}_t) < \theta$ **do**
$t = t + \delta$
5: **end while**
6: $\mathcal{S} \Leftarrow [t_0, t] \cup \mathcal{S}$
7: $t_0 \Leftarrow t$
8: go to line 3

The segmentation result depends on multiple parameters that need to be optimized:

- Depending on **the duration T of the sliding window**, there might be a delay before a scene boundary is detected. To get rid of this dependency, any boundary detected somewhere during a speech segment is moved to the beginning or the end of this segment (whichever is the closest).
- The **sliding window step** δ is arbitrarily set to 500 ms in this paper.
- A lower value for **threshold** θ tends to generate a larger number of segments.

Speaker-weighted distance d **–** It is obvious that some characters play a more important part than others in most TV series. Characters that only appear sparsely during an episode can be considered as minor characters (as opposed to recurring main characters). Therefore, we propose to take this difference into account by defining a speaker-weighted distance $d = d_\alpha$ as follows:

$$d_\alpha(\mathbf{x}, \mathbf{y}) = \frac{1}{D} \sum_{i=1}^{D} \alpha_i \cdot |x_i - y_i| \quad \text{where} \quad \sum_{i=1}^{D} \alpha_i = 1$$

α_i can be computed in several ways and depends on the total speech duration $\mathbf{L}(i)$ of the speaker i:

- $\alpha^=$ / same weight for all characters / $\alpha_i = \frac{1}{D}$
- α^+ / main characters weight more / $\alpha_i = \frac{\mathbf{L}(i)}{\sum_{j=1}^{D} \mathbf{L}(j)}$
- α^- / main characters weight less / $\alpha_i = 1 - \frac{\mathbf{L}(i)}{\sum_{j=1}^{D} \mathbf{L}(j)}$

3. COLOR-BASED SEGMENTATION

Our definition of a scene based on spatio-temporal continuity usually implies that video frames extracted from the same scene are visually similar.

Therefore, we choose to implement the method proposed by Yeung et al. [2] that relies on this characteristic: a scene is a succession of shots showing some kind of visual coherency. This approach is quickly described in Figure 3.

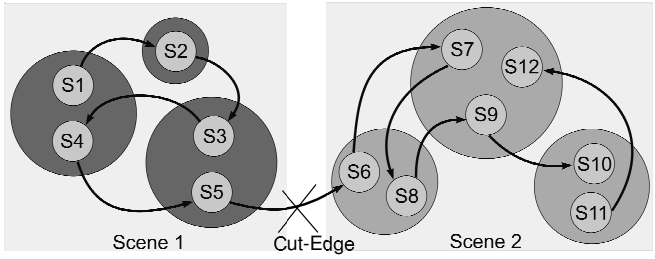

Fig. 3. Scene boundaries detection. S1 to S12 are the video shots. First, visually similar (and temporally close) shots are grouped together to form a collection of clusters (six, in this example). Then, using clusters as nodes, a graph is generated by linking all pairs of clusters containing temporally adjacent shots. Finally, cut-edges are removed, resulting in multiple disconnected sub-graphs: scenes.

4. AUDIOVISUAL FUSION

In order to achieve better segmentation results, we propose to combine the output of the audio-only system based on speaker diarization with the segmentation resulting from the visual-only color-based approach. From our various experiments described in the following paragraphs, we found out that the major issue with the speaker-based segmentation is that it does not take into account the actual video shot boundaries (on which groundtruth scene boundaries are aligned). It is therefore virtually impossible for such an approach to detect boundaries at their exact position, while the color-based segmentation is (by design) aligned on shot boundaries.

Consequently, our audiovisual fusion system consists in moving every audio scene boundary onto the closest visual scene boundary – and use the resulting modified speaker-based segmentation as the final audiovisual segmentation.

We introduce two ways of performing this fusion. The first one, denoted F, is the fusion of the best audio-only segmentation with the best color-based segmentation (parameters used for the speaker and color-based segmentation are learned separately). The second one, denoted F^*, consists in jointly optimizing the audio and visual parameters, with respect to the performance of the global audiovisual segmentation system.

5. EXPERIMENTS

5.1. Corpus

In order to perform our experiments, we acquired the first season of the *Ally McBeal* TV series. We manually annotated the first four episodes with shot and scene boundaries – for a total duration of around 3 hours of videos, 2788 shots and 239 scenes. We also annotated the four episodes with speaker segments, in order to evaluate the influence of the potential errors produced by the automatic speaker diarization system.

The whole set of annotations is made freely available on the Internet[1]. We also provide MFCC coefficients and HSV histograms extracted from the videos.

5.2. Evaluation metric

We consider the segmentation problem as a boundary detection problem and therefore rely on the well-known precision, recall and F-measure. The correctness of a boundary between scenes is defined in two different ways, depending on whether the evaluated approach is speaker- or color-based.

5.2.1. Evaluation of speaker-based segmentation

As highlighted in Section 4, it is very unlikely for an audio-only speaker-based segmentation system to detect the exact location of scene boundaries (which are aligned on visual shots, by construction).

This approach has no clue on how to decide on the actual position of a scene boundary detected during a non-speech segment. For instance, in Figure 4, there is no way for the audio-only system to decide on whether the second detected boundary is more relevant than the third one (as they both fall in the same non-speech segment).

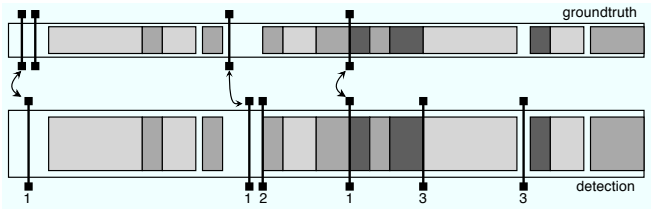

Fig. 4. Evaluation of speaker-based segmentation

Therefore, for evaluation purposes, a detected boundary is considered correct (marked with a **1** in Figure 4) if it is the first one detected in the same non-speech segment as the groundtruth boundary. All other detected boundaries in the same non-speech segment (marked with a **2**) are considered incorrect. A detected boundary is also considered incorrect (marked with a **3**) if no groundtruth boundary happens during the same non-speech segment.

5.2.2. Evaluation of visual and audiovisual segmentations

Since these segmentations output boundaries selected among shot boundaries[2], we consider a boundary to be correct if it has the exact same position as a groundtruth boundary (and incorrect otherwise).

5.3. Protocol

Since only four episodes are annotated, the evaluation protocol follows the *leave-one-out* cross-validation paradigm. Optimal parameters are obtained automatically by tuning the segmentation algorithms using three episodes (training set) and are applied on the remaining episode (validation set) to obtain the desired metric – this process being repeated for each episode. The final metric value is computed as the average of values obtained from the four combinations.

5.4. Results

Table 1 shows the results for our four segmentation systems. Fusion F only brings a tiny improvement over the color-based approach. However, fusion F^* shows that jointly training audio and video segmentations lead to an increase of the F-measure of nearly $+15\%$ compared to the color-based segmentation and even $+9\%$ compare to the speaker-based segmentation which is evaluated using a much more permissive protocol.

[1] `limsi.fr/Individu/bredin/publications/resources`

[2] The whole paper assumes that the list of shot boundaries is available.

Weights	speaker	color	Fusion	
			F	F^*
$\alpha^=$	0.317		0.312	0.341
α^+	0.297	0.309	0.311	**0.355**
α^-	**0.325**		**0.315**	0.350

Table 1. F-measure for speaker-based segmentation, color-based segmentation and their audiovisual fusion. Speaker-based systems shall not be compared to other approaches as they have a dedicated evaluation protocol (see Section 5.2).

	speaker	color	final
# boundaries	954	461	317
Precision	0.178	0.256	0.310
Recall	0.691	0.533	0.449
F-measure	0.270	0.331	**0.355**

Table 2. Insights into the best audiovisual system F^*

Table 2 allows for a better understanding of the fusion method F^*. It shows that both the audio and video approaches selected for the fusion tend to over-segment the videos: they detect 954 and 461 boundaries respectively, while the corpus only contains 239 scenes. Aligning the audio-only boundaries onto the closest visual ones allows to greatly reduce this undesired behavior (from 954 to 317 boundaries). Based on the observation of the improvement in terms of precision, it appears that most of the boundaries that are removed during the fusion process are actually incorrect boundaries.

We also underline that the F-measure values provided in Table 1 and Table 2 for the color-based segmentation and the various fusion approaches were obtained without allowing any temporal tolerance on the boundary location. In [5], the authors consider a boundary to be correct if it is within four shots from the groundtruth boundary – that is approximately 15 seconds in our corpus. Figure 5 shows that, under these circumstances, our proposed approach F^* reaches a F-measure of 0.725.

Finally, we observe that the color-based segmentation part of the fusion F^* (second column of Table 2) shows a better F-measure than the (supposedly) best color-only segmentation in Table 1. This observation uncovers the inefficiency of the current way of selecting the optimal parameters (i.e. grid search in a leave-one-out paradigm).

6. CONCLUSION

Through a novel approach based on the fusion of audio and video segmentations, we show that scene boundaries can be detected in TV series using speaker diarization.

Yet, there is still lots of room for improvement. For instance, we find that the optimal set of parameters vary a lot from one episode to another one. However, the training phase used in the current version of the algorithm prevents us from defining episode-specific parameters. One solution could be to introduce some kind of adaptive threshold θ or a new weighting scheme dependent on a local number of speakers, for instance.

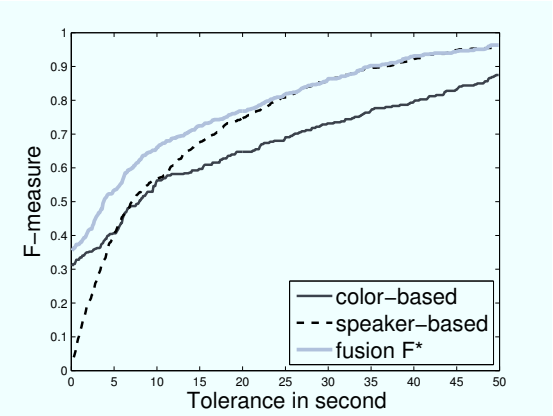

Fig. 5. F-measure as a function of the temporal tolerance

Finally, comparison with other scene segmentation tools is quite impossible due to the variety of content sets and evaluation protocols. To our knowledge there is no framework freely available today which would allow this comparison. So, by making the corpus annotations and descriptors freely available on the Internet, we hope it will encourage other researchers to publish results that can be easily and fairly compared.

7. REFERENCES

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

[2] M. Yeung, B. Yeo, and B. Liu, "Segmentation of Video by Clustering and Graph Analysis," *Comput. Vis. Image Underst.*, vol. 71, pp. 94–109, July 1998.

[3] L. Xie, P. Xu, S. Chang, A. Divakaran, and H. Sun, "Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models," *Pattern Recogn. Lett.*, vol. 25, pp. 767–775, May 2004.

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

[5] S. Zhu and Y. Liu, "Video Scene Segmentation and Semantic Representation Using a Novel Scheme," *Multimedia Tools Appl.*, vol. 42, pp. 183–205, April 2009.

[6] E. El Khoury, C. Senac, and R. Andre-Obrecht, "Speaker Diarization: Towards a More Robust and Portable System," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 2007, vol. 4, pp. 489–492.