From Video and Audio Recurrences to Unsupervised Program Structuring

Alina Elma Abduraman and Sid-Ahmed Berrani
Orange Labs – France Telecom R&D
35510 Cesson-Sévigné. France.
{*alinaelma.abduraman, sidahmed.berrani*}*@orange-ftgroup.com*

Bernard Merialdo
Multimedia Communications Dept. – Eurecom
06904 Sophia Antipolis. France.
Bernard.Merialdo@eurecom.fr

***Abstract*—This paper addresses the problem of unsupervised TV programs structuring. Program structuring allows direct and non linear access to the desired parts of a program. Our work addresses the structuring of recurrent TV programs like news, entertainment programs, TV shows, TV magazines... In our previous work [1] we proposed a program structuring method based on the detection of video recurrences. In this paper we extend our study to audio recurrences and verify their influence on the final structuring. We evaluate the structuring results on both approaches (audio and video) separately and jointly. We use for evaluation a 62 hours dataset corresponding to 97 episodes of TV programs.**

***Keywords*-TV Program Structuring, Non-linear Browsing, Audio and Video Recurrence Detection, TV Content Indexing.**

I. Introduction

Automatic video segmentation and indexing provide access to meaningful/relevant parts of a video. The idea is to extract features that allow users to directly and quickly find interesting moments in a video. This processing step cannot be performed manually when dealing with the huge amount of available video content, in particular when dealing with broadcasted TV content.

This paper focuses on this content, and more specifically on TV program structuring. TV programs have an underlying structure that is generally lost when programs are broadcasted. The only available reading mode when viewing TV programs that are recorded using a Personal Video Recorder or through a TV-on-Demand service is the linear mode. Thus, when one would like to skip a part of the program and to go to another part, he/she has the unique and not very practical mean: the fast-forward function.

In this context, program structuring becomes thus important in order to provide users with a novel and useful browsing features. Basically, program structuring consists in recovering the original structure of the program. It enables finding the start time of each part composing the program.

In addition to advanced browsing features, TV program structuring can also be used for summarization. Recovering the structure allows to build a balanced summary where each part of the program is represented with respect to its importance. Indexing and querying, commercial skipping, archiving, intra-program audience measurement are also possible applications.

Due the absence of metadata on the structure of programs, program structuring mainly relies on the analysis of the audio-visual signal of the TV program. One commonly used feature in most of the recent works in the field is "the recurrence" [1], [2], [3], [4]. Indeed, it is very frequent for programs to be composed of recurrent segments that act as anchor points in programs. Examples of such recurrences are anchor person shots in news, audio jingles that announce the passing from a topic to another in a TV/radio magazine or the passing to another stage of a TV game show. These recurrences are introduced on purpose in order to allow viewers/listeners to easily follow the program and identify its structure even if they do not watch it from the beginning.

In our previous work [1], we have considered only visual recurrences. In this paper, we propose to extend the study by considering also audio recurrences. We propose an audio recurrence detection method and we evaluate its ability to improve the quality of the final structuring results when combined with the video-based recurrence detection method.

We focus in this work on *"recurrent"* TV programs. A recurrent TV program is a program composed of several "episodes" that are periodically broadcasted (i.e. daily, weakly, monthly...). Examples of this type of programs are entertainment programs, game shows, magazines, news... We mainly explain our choice for this type of programs by the applicative interest this type of programs have as they represent about 50% of the total number of broadcasted programs of the French generalist TV channels[1]. Moreover they have important properties that make our task feasible, meaning they generally have a clear structure with well defined main parts. These are generally delimited by short video/audio sequences that we will call *"separators"*. These are recurrent sequences as they repeat within or between different episodes of a TV program. This allows us to detect the separators as video/audio recurrences between/within several episodes of the same recurrent TV program, in an unsupervised manner. Using separators, the different parts of the TV program can be delimited. Consequently, there is no need to any prior knowledge on the structure or on the number of parts that the program might have.

The rest of the paper is organized as follows. Section II

[1]Source : Mediamat-Mediametrie, 2010

presents existing techniques for program structuring. Section III presents the proposed approaches for audio and video separators detection. Section IV evaluates the contribution of audio and video recurrences in the detection of separators, concludes the paper and discusses future extensions.

II. Related work

Program structuring has been extensively studied. A first class of approaches tries to structure programs in a supervised manner exploiting the prior knowledge of the domain in order to extract relevant data and construct a structured model of the analyzed video. They generally require the prior creation and manual annotation of a training set that is used to learn the structure. This approach has been widely applied to TV news videos. Template matching techniques based on weighted similarity measures, trained models and predefined rules are generally used [5], [6].

Sport programs are also usually analyzed in a supervised manner. Rules of the games provide prior knowledge that can be used to provide constraints on the appearance of events or their succession. The idea is to segment in a first step the video into narrative segments (like play and break) making use of the production rules and the different outputs of camera views [7] or based on predefined models [8]. Events are then detected based on object tracking (e.g. player tracking [9] and localization [10], ball tracking [11], detection of goal mouth, center line [12], red/yellow cards [13]) and/or specific sounds identification (exciting speech, applause, hitting ball [14]...).

All these approaches are however supervised as they rely on the prior knowledge of the domain. They depend on the genre of the program and its production rules.

The challenge becomes thus the development of content-based automatic tools. The objective is to allow the easy and unsupervised production of structure information.

A lot of literature has developed on the detection of scenes in videos as structuring elements. The proposed approaches use clustering methods [15], scene transition graphs [16], cinematic rules [17], HMMs [18] or SVM classifiers [19]. However, the definition of a scene is very ambiguous and depends on the subjective human understanding of its meaning. Consequently, these methods are difficult to evaluate.

Another category of very recent approaches are those based on the recurrence of certain audio/video sequences. The recurrences are structuring elements and are used for recovering the underlying structure of a program. In [4] the video self-similarity is exploited to identify the frequent patterns occurring in the video signal. In [1], separators inserted between different parts of a program are detected using a video repeated sequence detection method applied over a set of episodes of the same recurrent TV program. Audio recurrences are also detected in radio broadcast streams in order to identify the jingles that separate different topics or to extract meaningful information that could serve for audio summarization [20]. In [20], a dynamic time wrapping technique is used to discover repeating words. The principle is inspired from [3] where audio repetitions in audio streams are detected by time correlating low-dimension audio representations. In this case, the video repetitions are used to validate the audio ones. Audio and video consistency of repeated segments is used in [2] where an event mining technique is proposed. A clustering method is applied for each modality separately. Then, a measure of audiovisual consistency between the clusters resulted from the two modalities is used to select the candidate events. Knowledge about the targeted event needs to be used in order to select relevant events for structure analysis, using an SVM classifier.

III. The proposed approach

We focus in this work on structuring recurrent TV programs. One of the main properties of these programs is their clear and steady structure. Moreover, the different parts of a program are generally delimited by short audio/video sequences that we named *"separators"*. In Figure 1, examples of separators are illustrated. In order to avoid any confusion, we propose to use the term *program* to refer to a recurrent TV program and *episodes* to refer to the different broadcasts of a recurrent TV program. The horizontal lines represent the timeline of different episodes of the same program. The boxes represent the separators that delimit the main parts of the program. The 3 images are extracted from three separators of a French game show.

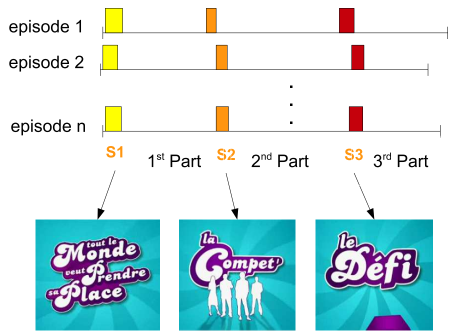

Figure 1. Example of video separators.

The main idea of our approach is to automatically detect these separators. The different parts of each episode are then identified using the separators boundaries, e.g. the end of a separator is the start of a new part. The resulting structure for an episode is hence composed of a set of timecodes that refer to the start and end times of each part.

To achieve this, our approach analyze a set of episodes of a recurrent program in order to detect the separators as recurrent sequences. It can also be used to structure a new episode by including it into a set of previously broadcasted episodes to be analyzed.

In our previous work [1], we have considered only video recurrences. We were then able to detect only separators that have the same visual content. Our previous solution did not address the audio jingles that are separators which share the same audio content but not necessarily the same visual content (i.e. the same sequence of images). In this paper, we propose to extend our previous approach by considering also audio recurrences. We propose hence a new audio recurrence detection method.

In the following subsections, we first recall our approach for video separator detection. We then describe the approach for audio separator detection.

A. Video separators detection

The unsupervised detection of the video separators makes use of two important properties the separators have:

1) Repeatability: Different episodes of the same recurrent TV program share the same structure and their separators are almost identical. This repeatability of separators can be found between episodes of the same recurrent TV program (inter-episode), but also inside a single episode (intra-episode).
2) Temporal stability: As the different parts of a program have approximatively the same duration, a separator can be found at approximately the same time-offset, for different episodes of the same recurrent program.

The 2 main steps of the video separators detection are described below.

1) Video recurrences detection: This step performs the description of the visual content for episodes that are going to be structured. It also proceeds with detecting the repeated video sequences using the technique described in [21]. It based on a clustering of the keyframes in order to group together the near-identical shots and to detect the set of video recurrences. For more details the reader can refer to [21].

2) Separator detection: All the occurrences of the recurrences, detected during the previous step, are not necessarily separators. It might happen that within the analyzed episodes there are sequences that are replayed or that are very similar. These of course are not separators. For instance, shots showing the moderator in the same position but at different times of the episode might be detected as occurrences of a repeated sequence. They are not separators but their content is very similar. We call these occurrences "*false alarms*". In order to remove them, a post-processing step is used.

The detected repeated sequences are first passed through a first filter that removes a repeated sequence that has all its occurrences only coming from the same episode. Even if a separator can be repeated within the same episode, it has to be also repeated over at least two episodes in order to be valid. It is very unlikely to have a separator created specifically for a single episode.

As explained previously, in the case of inter-episode repeated sequences, false alarms may also appear. In order to remove them, the second property of the separators is exploited. To achieve that, a study of the temporal density of the occurrences of detected recurrences from the input episodes is performed. All the occurrences from different episodes are projected on the same temporal axis. From this projection, a histogram is computed by counting the number of occurrences during each 40ms window (each frame). A kernel-based density estimation is then performed. In this study, a Gaussian kernel has been used:

$$f_i = \sum_{j=i-3\sigma}^{i+3\sigma} h_j e^{-\frac{(j-i)^2}{2\sigma^2}}, \tag{1}$$

where f_i represents the filtering result for frame i and $h(j)$ represents the number of occurrences computed from the histogram, corresponding to frame j.

The idea is to find the areas with high concentrations. These are likely to be times when a separator is broadcasted. Isolated occurrences are likely to be false alarms as they appear quasi-randomly without any temporal stability.

The result of the temporal density analysis is a distribution curve where a maximum represents an area of high concentration of the separators. A threshold is then defined as a fraction of the mean of all the maxima. The separators that have a density under the threshold are rejected. An illustrative example is given in Figure 2. The occurrences in dark are isolated occurrences.

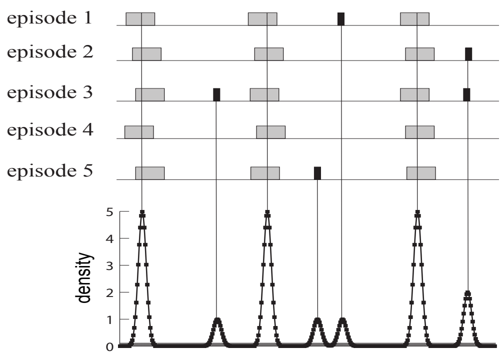

Figure 2. Temporal density of detected occurrences.

B. Audio separators detection

For the audio separators detection we follow the same idea as for the video ones. In order to detect the audio recurrences, we apply a recurrence detection algorithm over a set of episodes of a recurrent TV program. Among the detected recurrences there are also the separators.

The audio recurrence detector is composed of the following steps:

Step 1: For each of the episodes in the analyzed set, a *descriptor* is computed. The *descriptor* contains a set of time-sorted *sub-descriptors* of the current episode and some identification data. The *sub-descriptors* are composed of basic audio features, extracted from each episode separately using a sliding audio frame. The descriptors computed for all the episodes in the analyzed set are stored in a database.

Step 2: When searching for the recurrences, each of the episodes, in turn, will become query while the rest will become references. A sliding window (having a predefined number of sub-descriptors) that we call *TimeSlot* is used to match the sub-descriptors of the query with those of the references. The distance chosen to compare the query and the reference is the average Hamming distance computed on all the sub-descriptors of the query and reference TimeSlots. A match is confirmed if the distance between the two falls below a pre-determined similarity threshold. In order to detect the intra episode recurrences as well, the query is also matched with itself. At the end of the processing, a set of matching segment pairs are obtained (see Figure 3).

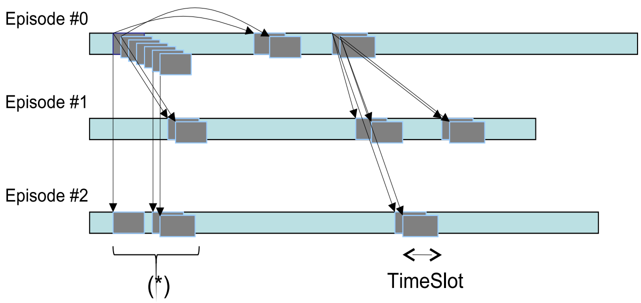

Figure 3. Audio segment recurrence detection.

Step 3: The identified detections in step 2 are smoothed and extended with overlapping segments. If the distance between two detected segments falls under a certain threshold, these segments are merged (see (*) for Episode #2 in Figure 3 and in Figure 4).

Step 4: The matched segments are analyzed and clustered. First, each pair of segments obtained in step 3 instantiates a cluster. Second, each two overlapping clusters are merged. Two clusters overlap if a segment from the same program belongs simultaneously to both clusters. Figure 4 shows the results of the clustering corresponding to the detections in Figure 3. The two colors of the segments in Figure 4 (black and gray) correspond to the two clusters to which the sets of matching segments in Figure 3 belong.

As already discussed in the second step of the video recurrence detection algorithm, not all the detected recurrences are necessarily separators. In this case too, we can apply the temporal filters described in the previous sub-section, in order to remove these *"false alarms"*.

IV. Experiments

To evaluate the influence of video and audio recurrences over the structuring of TV programs, we have performed experiments using real TV broadcasts. We used a variety of

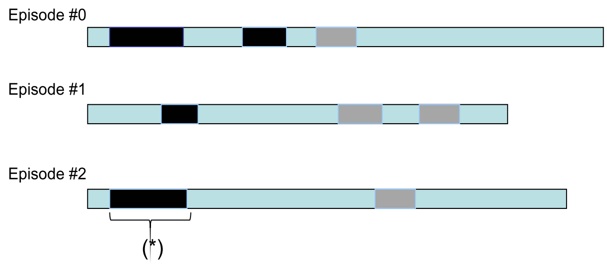

Figure 4. Clusters of matched pairs of audio segments.

TV programs, excluding movies and TV series which are not our concern. We discuss the results obtained using the 2 approaches separately but also when combining them.

A. Datasets

Our experimental dataset contains of 62 hours of videos, corresponding to 97 episodes of TV programs broadcasted on four French TV channels (TF1, France2, M6, Orange).

Two of them, denoted *Motus* and *Tout le monde veut prendre sa place*, are TV games. The separators in this case delimit the different stages of the game. The episodes, have only inter-episode repeated separators, i.e. there is no separator that is repeated inside the same episode. The duration of an episode varies between 30 and 45 minutes.

Two others denoted *10h Le Mag* and *Orange Cinema Series*, are TV magazines. The separators in this case delimit the reports and the scenes where the anchor presents the next topic. They are both inter and intra-episode repeated separators. The episodes last about 50 minutes.

The last program is a news program and it is similar to the two previous ones. It lasts 20 minutes and it is composed of a set of reports and anchor person scenes that are separated by inter and intra episodes repeated separators.

In order to study both intra- and inter-episode recurrence, for each TV program, we divided the dataset into subsets. Each subset is composed of two episodes. We performed experiments on each subset separately. In this manner the results will contain both intra- and inter-episode recurrences.

For all the episodes of these programs, we have manually created a ground-truth. Each video separator has been precisely determined, indicating its start time and end time. As we use the same dataset for the case of video and audio approaches, the audio and video separators will coincide. We use thus the same ground truth for both video and audio. Table I provides a description of these datasets.

All the separators are composed of images combined with audio like music, recorded/natural speech or specific sounds. Our interest is to show the influence of each type of recurrence (audio/video) over the detection of separators and thus over the segmentation results. We are also interested in evaluating the improvement that could be brought or not in the detection of separators, when combining the

	Dataset name	No. of episodes	No. of separators
Motus	M	26	4
TLMVPSP	T	31	5
10H Le Mag	L	10	20-23
Orange Cinema Series	C	16	15-17
20h45	N	14	20-27

Table I
DATASETS DESCRIPTION.

two approaches. Consequently, we have used as evaluation criteria the recall measure, computed only on the second episode of each subset (as it would be the case in a real world service). We will not focus in this paper on the filtering step.

A separator is considered as being correctly identified if it overlaps with its correspondent in the ground truth by at least a predefined threshold (ς).

B. Video and audio separators evaluation

In order to validate a video separator we impose a threshold of ς = 50%. For the evaluation of audio results the ground truth created for the video separators was used. Consequently, we show also the results obtained using a smaller ς meaning, 30%, as it may happen for the audio separator and its video correspondent not to coincide perfectly. The obtained results are presented in Table II.

Datasets	Video Recall	Audio Recall		Audio-Video Recall	
	ς=50%	ς=50%	ς=30%	ς=50%	ς=30%
M	0.82	0.46	0.47	0.83	0.83
T	0.70	0.53	0.62	0.80	0.84
L	0.66	0.92	0.96	0.98	0.98
C	0.07	0.66	0.80	0.66	0.80
N	0.06	0.50	0.85	0.50	0.87

Table II
RECALL OF VIDEO AND AUDIO SEPARATORS.

Effectiveness of video recurrence detection

The second column of the Table II presents the recall for the video separators alone. For the game shows, the results show a good detection of separators with high recall values. This is explained by the fact that in the case of games, the separators are identical sequences of synthetic images that can be easily recognized by our algorithm. The games contain some sequences that present an object that will be won. This object may or may not be different from one show to another. Hence, the repetition based algorithm will or will not be able to detect the corresponding sequence. Including them or not in the ground truth may change the results. For instance, for the case of dataset M, when not including these sequences in the ground truth, the recall increases to 0.97.

For datasets L, the separators are also composed of synthetic images that do not always succeed in the same order and that vary in length. This is why often only small parts of them are identified by the recurrence detection algorithm. Decreasing the ς parameter to 30% will increase thus the recall. Also, applying the algorithm on only two episodes, we risk to miss the detection of separators (that are sligthly different from those in the current episodes) that are not repeated within the considered episodes [2].

For the last two datasets our approach is not fitted for the specific type of separators that occur. The separators for the news program (N) are composed of a natural image with the anchor person (in different positions and with different backgrounds) on which a logo is superposed. For the case of the magazine C most of the separators are composed of synthetic images that are sometimes slightly different or that have different editing effects. As our approach is conceived to detect nearly identical sequences most of these separators will not be identified.

Effectiveness of audio recurrence detection

Colums 3 and 4 of the Table II illustrate the results obtained when using the audio approach alone.

For the case of TV games, the audio recurrence detection algorithm does not perform as well as the video one while for the other datasets, the algorithm performs better. The separators that have not been detected are generally separators containing natural speech. Even if the person who speaks says the same thing, she/he can say it with different frequencies. As the audio descriptors we used in our approach are not suited for spoken recurrences, these separators will not be detected resulting thus in smaller values for the recall. These separators have been however detected by the video based approach as the sequence of images from the separator remains the same. This is moreover the case for datasets M and T. Also, separators that have a noise on the background (like applause or speech) are detected only if during the clustering, they are compared to other similar but clean separators. If two identical sequences but that have both applause on the background are compared they will not be found as similar. This problem appears for some separators in datasets L and C. For dataset N, which refers to the news program, there are some separators that are only visual. They mark the transition from the anchor person scenes, where the next topic is introduced, to the report that presents this topic. They are composed of only images with a map of the country where the report was filmed. If more reports from a same country are presented our video approach identifies the corresponding separators.

Combining audio and video results - Conclusion

The last two columns of the Table II join together the results of the video and audio approaches. We obtained them by making the union between the results obtained for each modality separately. As it can be seen, using both audio

[2] for more details see [1] where experiments were made on complete weaks of episodes, for some of the datasets in our current evaluation

and video approaches can, more or less, improve the recall. This depends on the type of programs and on the type of separators these programs have.

Consequently, we can say that both audio and video approaches should be used for the detection of separators and the structuring of episodes. Using both of them, a large category of TV programs (and in consequence a high variety of separators) could be treated. As noticed earlier in this section, both approaches have their limitations. For instance in the case of audio recurrence detection algorithm the separators containing natural speech can not be detected. For the case of the video recurrence detection algorithm, the separators that are not identical are not detected. Or, it might happen that the clustering algorithm, which involves an optimal setting of large number of parameters, to wrongly perform for certain specific cases. When using both approaches, one could overcome the limitations of the other and provide thus the necessary for a better structuring.

Our future work will focus on the study of models that could integrate the two proposed approaches. We also think that it would be interesting to classify the separators into different categories. This will allow an easier analysis of the results. An important point will also be to relate the recall and precision measures to the end-user satisfaction.

REFERENCES

[1] A. E. Abduraman, S.-A. Berrani, and B. Merialdo, "An unsupervised approach for recurrent tv program structuring," in *9th edition of the European Interactive TV Conference*, Lisbon, Portugal, June-July 2011, pp. 123–126.

[2] M. Ben and G. Gravier, "Unsupervised mining of audio-visually consistent segments in videos with application to structure analysis," in *IEEE International Conference on Multimedia and Exhibition*, Barcelone, Espagne, July 2011, pp. 1–6.

[3] C. Herley, "Argos: automatically extracting repeating objects from multimedia streams," in *IEEE Transactions on Multimedia*, vol. 8, no. 1, 2006, p. 115129.

[4] A. Jacobs, "Using self-similarity matrices for structure mining on news video," *Advances in Artificial Intelligence*, vol. 3955, pp. 87–94, 2006.

[5] B. Gunsel, A. Ferman, and A. Tekalp, "Temporal video segmentation using unsupervised clustering and semantic object tracking," *Journal of Electronic Imaging*, vol. 7, no. 3, pp. 592–604, 1998.

[6] H. Misra, F. Hopfgartner, A. Goyal, P. Punitha, and J. M. Jose, "Tv news story segmentation based on semantic coherence and content similarity," in *16th International Multimedia Modeling Conference*, Chongqing, China, January 2010, pp. 347–357.

[7] D. W. Tjondronegoro and Y.-P. P. Chen, "Knowledge-discounted event detection in sports video," *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 40, no. 5, pp. 1009–1024, September 2010.

[8] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden markov models," *Pattern Recognition Letters*, vol. 25, no. 7, pp. 767 – 775, May 2004.

[9] K. Kim, M. Grundmann, A. Shamir, I. Matthews, and J. H. I. Essa, "Motion fields to predict play evolution in dynamic sport scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June 2010, pp. 840 – 847.

[10] R. Hamid, R. K. Kumary, M. Grundmannz, K. Kimz, I. Essaz, and J. Hodgins, "Player localization using multiple static cameras for sports visualization," in *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June 2010, pp. 731 – 738.

[11] A. Guziec, "Tracking pitches for broadcast television," *Computer*, vol. 35, no. 3, pp. 38 – 43, March 2002.

[12] X. Yu, L. Li, and H. W. Leong, "Interactive broadcast services for live soccer video based on instant semantics acquisition," *Visual Communication and Image Representation*, vol. 20, no. 2, pp. 117–130, February 2009.

[13] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796–807, July 2003.

[14] L.-Y. Duan, M. Xu, T.-S. Chua, Q. Tian, and C.-S. Xu, "A mid-level representation framework for semantic sports video analysis," in *ACM International Conference on Multimedia*, Berkeley, CA, USA, November 2003, pp. 33 – 44.

[15] J. Wang, L. Duan, Q. Liu, H. Lu, and J. S. Jin, "A multimodal scheme for program segmentation and representation in broadcast video streams," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 393 – 408, April 2008.

[16] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso, "On the use of audio events for improving video scene segmentation," in *11th International Workshop on Image Analysis for Multimedia Interactive Services*, Desenzano del garda, Italy, April 2010, pp. 1–4.

[17] J. Wang and T.-S. Chua, "A cinematic-based framework for scene boundary detection in video," *The Visual Computer*, vol. 19, no. 5, pp. 329–341, 2003.

[18] Y. Zhai and M. Shah, "Video scene segmentation using markov chain monte carlo," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 686 – 697, August 2006.

[19] N. Goela, K. Wilson, F. Niu, and A. Divakaran, "An svm framework for genre-independent scene change detection," in *IEEE International Conference on Multimedia and Expo*, Beijing, China, July 2007, pp. 532–535.

[20] A. Muscariello, G. Gravier, and F. Bimbot, "Audio keyword extraction by unsupervised word discovery," in *Conference of the International Speech Communication Association (Interspeech)*, Brighton UK, 2009, pp. 2843–2846.

[21] S.-A. Berrani, G. Manson, and P. Lechat, "A non-supervised approach for repeated sequence detection in tv broadcast streams," *Image Communication*, vol. 23, no. 7, pp. 525–537, August 2008.