Unsupervised mining of multiple audiovisually consistent clusters for video structure analysis

Anh-Phuong Ta, Guillaume Gravier

UNSUPERVISED MINING OF MULTIPLE AUDIOVISUALLY CONSISTENT CLUSTERS FOR VIDEO STRUCTURE ANALYSIS

Anh-Phuong TA[1] *and Guillaume Gravier*[2]

[1]INRIA-Rennes, Campus Beaulieu, F-35042 Rennes, Cedex, France. Email: *anh-phuong.ta@inria.fr*
[2]CNRS-IRISA, Campus Beaulieu, F-35042 Rennes, Cedex, France. Email: *guillaume.gravier@irisa.fr*

ABSTRACT

We address the problem of detecting multiple audiovisual events related to the edit structure of a video by incorporating an unsupervised cluster analysis technique into a cluster selection method designed to measure coherence between audio and visual segments. First, mutual information measure is used to select audio-visually consistent clusters from two dendrograms representing hierarchical clustering results respectively for the audio and visual modalities. A cluster analysis technique is then applied to define events from the audio-visual (AV) clusters with segments co-occurring frequently. Candidate events are then characterized by groups of AV clusters from which models are built by automatically selecting positive and negative examples. Experiments on the standard Canal9 data set demonstrates that our method is capable of discovering multiple audiovisual events in a totally unsupervised manner.

Index Terms— Multiple events, Video mining, Video structuring, Cluster selection, Mutual Information, Event discovery, Structural event, Audiovisual consistency.

1. INTRODUCTION

Generally speaking, video structuring consists in extracting semantic events—e.g., actions in sport videos, violent scenes in movies, etc.—and/or events related to the edition of the video—e.g., monochrome frames, dissolve, shot boundaries, etc.—so as to segment the video into its constituents. In this work, we define a structural element as a key content that appears frequently in a video and exhibits audio and visual consistency. Typical examples of such events are jingles in news videos, anchor persons or participants in a talk show, etc. Existing approaches to video structure analysis fall into two main categories: i) segmentation of the entire video (referred as *dense* segmentation), where the input video is mapped to a predefined structure, like points, games or sets in tennis videos; ii) detection of specific events related to the video structure, such as advertisements or goals in sport videos. This paper presents a novel approach belonging to this last category where events are defined from the data rather than a priori. We focus on the detection of multiple events relevant to the structure of a video, in a totally unsupervised fashion without any prior knowledge about the events to be detected. Our aim is to propose a generic method for video structure analysis, which can have applications in semi-supervised video annotation and edition, automatic structuring of videos, summarization, etc.

In the multimodal video mining literature, many efforts have been focusing on supervised learning (see [1]) and content-based analysis techniques such as speech recognition or face detection and identification. For instance, such approaches have been employed for anchor person detection [2], or for the detection of specific events like goals in sport videos [3]. Similarly, Li *et al.* [4] proposed to combine face recognition and speaker detection to find occurrences of characters in movies. Despite their success, the common limitations of such supervised methods are the need to train a model from manually annotated data and the lack of robustness to unseen data. Moreover, some methods require a manual initialization step, e.g., [4], therefore, lacking the generality to cope with diverse video genres.

As an alternative, detecting repeating patterns has been considered [5][6][7]. However, such methods focus on discovering near-duplicate repetitions, and cannot deal with variations across repetitions (i.e., the repetitions are not exact), which is a crucial issue in video structuring. The problem of mining repeating structural elements has also been addressed using clustering techniques [8][9][10][11] to group video shots exhibiting a strong visual similarity, which are likely to be relevant with respect to the structure of the video. However, clustering-based techniques cannot avoid the non-trivial problems of choosing the optimal number of clusters, and of dealing with outliers (as most of the data does not fit into any cluster).

To overcome these drawbacks, we propose an unsupervised approach to detect multiple events exhibiting a strong audio and visual consistency, often related to the video editing. We elaborate on the work of Ben and Gravier [12][13] and of Dielmann [11]. The former have proposed an unsupervised method to detect a single audiovisual structural event without any prior knowledge. From two dendrograms representing hierarchical clustering results of the audio and visual modalities, they measure the consistency between an audio

cluster and a visual cluster using mutual information in the temporal domain. Several heuristics are then applied to select a unique pair, made of an audio and a visual clusters, relevant to the video structure. As in most cases, discovery of multiple events is not considered while many videos exhibit several structural events (e.g., two anchor persons, guests in talk shows). However, the work of Dielmann [11] was designed to select multiple pairs of audio and visual clusters from two independent partitions of the data. To this end, Pearson's χ^2 statistical test is adopted to analyze the co-occurrences between audio and visual labels (clusters). AV-clusters are then identified as the ones whose labels most frequently co-occur. However, this method entirely relies on the performance of the partitioning algorithms used to construct two sets of labels corresponding to, resp., the audio and visual modalities. In particular, the number of clusters in each partition has to be defined in some way.

In this paper[1], we introduce a new scheme which combines the use of nested clusters and mutual information criterion with the χ^2 statistical test to select multiple pairs of audio-visual clusters, thus defining several structural events from the data. Note that, this paper differs from [11] in two major ways. First of all, we propose a method that does not need initial partitioning of the audio and visual data. Moreover, the absence of labels attached to segments as a result of the partioning steps requires significant adaptation of the χ^2 criterion. In particular, as opposed to the detection of only a single event in [12], we propose a fully unsupervised method able to detect multiple sets—clusters—gathering audiovisually consistent segments, selecting and grouping pairs of audio and visual clusters from two independent hierarchical clustering trees, one for each modality. Further, automatic selection of positive and negative samples enables to refine the results using support vector machines in addition to the initial discovery step.

The rest of the paper is structured as follows. Section 2 presents the general framework to discover multiple events. In section 3, we briefly present the audio and visual clustering step. We discuss how to select structurally consistent clusters in section 4, and presents our adaptation of the χ^2 test for relevant event identification in section 5. Event characterization and modeling is discussed in section 6. In section 7, we describe experimental results, followed by the conclusion and future work in section 8.

2. OVERVIEW

The general idea of the algorithm, illustrated in Figure 1 is as follows: two segmentations are built independently for the audio and visual modalities and a set of nested clusters is established for each modality using hierarchical bottom-up clustering. This initial step results in two independently constructed dendrograms, where each node represents a set of (supposedly) coherent segments. We then explore cross-modal relations to select pairs of audiovisual clusters—one from each modality—which define consistent audiovisual segments. This idea is introduced in [12] where the single most consistent pair of clusters is selected according to some heuristics on the pattern of occurrence of structurally relevant events. A rather similar philosophy is used in [11] to select pairs of segments assuming each segment is labeled. The key difference is that in [11] a unique segmentation is used in each modality, with cluster labels attached to segments, rather than a nested hierarchy of clusters. Exploiting the general idea of selecting consistent pairs of clusters, we investigate the selection of multiple events, combining several criteria. First, a list of candidate pairs is constructed from the N most consistent AV-cluster pairs according to the mutual information criterion. These candidates are then filtered using a χ^2 test and events are defined from the filtered list by grouping pairs corresponding to the same underlying event. Finally, for each candidate group (or event), segments are automatically selected to train a SVM classifier which is used to refine event detection. In the next sections, we discuss, in turn, each step of the process.

3. SEGMENTATION AND HIERARCHICAL CLUSTERING

The audio and video streams are first independently segmented into audio and video segments, respectively. For each modality, a classical bottom-up clustering technique is used to create a set of nested clusters represented as a dendrogram. The dendrogram encodes the various stages of the hierarchical clustering and each node in the dendrogram corresponds to a set of segments, either in the video or in the soundtrack. Audio segmentation implements a standard Bayesian information criterion to detect abrupt changes in the signal. Gaussian mixture models (GMM) are used to model each segment and an approximation of the Kullback-Liebler divergence between two GMMs is used for agglomerative bottom-up clustering. This approach, commonly used in speaker segmentation systems, groups segments with similar audio contents, e.g., sharing the same type of music or the same speaker voice. The video is segmented into shots based on color histograms to detect changes across different frames and a keyframe is extracted for each shot. Keyframe clustering is also color-based, each shot being represented by its color histogram in the RGB space with 8 bins per color. Euclidian distance and Ward's linkage are used in bottom-up clustering.

4. CONSISTENT CLUSTER SELECTION

Given the audio and visual dendrograms, the next step consists in selecting relevant pairs of audiovisual clusters (AV cluster pair), where an AV cluster pair consists of one node

[1]This work was partly funded by OSEO, French State agency for innovation, in the framework of the Quaero research program.

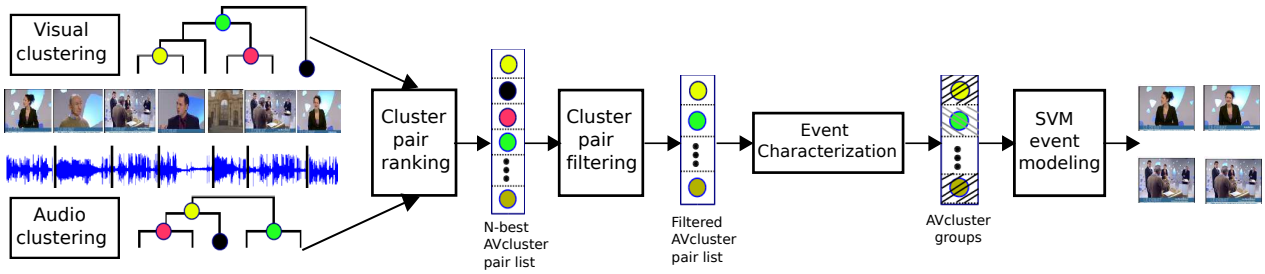

Fig. 1: Schematic illustration of our approach to detecting multiple events (best viewed in color).

from the audio dendrogram and one from the video one. Hence, an AV cluster pair is defined by the set of audio segments corresponding to the audio cluster and the set of video segments which corresponds to the visual cluster.

Let (C_i^A, C_j^V) be an AV cluster pair composed of the i-th and the j-th clusters of the audio and video dendrograms, respectively. Our objective is to measure the consistency between the AV cluster pair. To this end, the mutual information (MI) is applied:

$$MI(C_i^A, C_j^V) = \sum_{(a,v)\in\{(0,0),(1,1)\}} p(a,v)\ln\left(\frac{p(a,v)}{p(a)p(v)}\right) \quad (1)$$

where a and v are binary random variables which indicate membership in C_i^A and C_j^V, respectively. The probabilities $p(a,v)$, $p(a)$, and $p(v)$ are estimated from the temporal segmentation. For example, the join probability $p(a=1, v=1)$ is measured as the amount of time that segments of C_i^A and segments of C_j^V co-occur, normalized by the total duration of the video. A large value of the MI therefore indicates that the two corresponding clusters are closely consistent with each other.

Audiovisual consistency is measured using Equation 1 for each possible cluster pair and a list of the N best pairs is established. It must be noted that some pairs in the N-best list are strongly overlapping due to the fact that a dendrogram defines a set of nested clusters. Moreover, some AV cluster pairs might be irrelevant for video structure analysis. The next two steps of the algorithm therefore consist in selecting relevant pairs before grouping overlapping pairs which correspond to the same underlying event in the video. We detail the first step in the next section.

5. RELEVANT EVENT IDENTIFICATION

In [12], the selection of a single AV cluster pair was considered using heuristics to measure the relevance to the structure of the video. However, this only allows for the selection of a single pair and is highly application dependent. Instead, inspired by the work from Dielmann [11], we use a cluster analysis technique based on Pearson's χ^2 test to identify relevant pairs in the N-best list generated using mutual information.

5.1. Cluster analysis using χ^2 test

Pearson's χ^2 test can be used to verify whether two random variables are statistically independent or not. Particularly, Dielmann used χ^2 to compare pairs of labels (clusters) resulting from an independant partitions of the audio and visual modalities[2]. The test intends to determine whether a pair of labels (clusters) resulting from the partitioning step, expressed in a contingency table (matrix), are independent of each other (i.e., the null hypothesis) or not. In other words, the test consists in determining if two labels jointly occur more frequently than at random. Formally, let O be a $U \times V$ matrix, where each entry O_{ij} represents the number of times that a pair (i,j) (i.e. the co-occurrence of an audio segment labeled i and a video segment labeled j) is observed, and U, V are the audio and visual dictionaries, respectively. Under the null hypothesis that the occurrence of a video cluster and an audio cluster is statistically independent, the maximum likelihood estimated probability $p(i,j)$ of the pair of labels (i,j) is given as

$$p(i,j) = p(i)p(j) = \frac{\sum_{k=1}^{U} O_{ik}}{N}\frac{\sum_{l=1}^{V} O_{lj}}{N} \quad , \quad (2)$$

where N is the total number of observations, i.e., the sum of all entries in the matrix O. Under the null hypothesis, the expected (theoretical) frequency for any pair of labels is given by $E_{ij} = N.p(i,j)$. The χ^2 aims at identifying audiovisual label pairs (i,j) which co-occur more frequently than the expected frequency under the null hypothesis E_{ij}. Globally, the value of the test statistic χ^2 is given by:

$$X^2 = \sum_{i=1}^{U}\sum_{j=1}^{V}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad (3)$$

The higher the value of X^2, the bigger the deviation from the expected value under the null hypothesis and, hence, the higher the confidence that the two partitions coincide.

[2]Here, a partition is a segmentation with arbitrary labels attached to segments. The audio partition is the result of a speaker diarization algorithm while shot clustering is considered to obtain arbitrary labels for video shots.

5.2. Identification of candidate events

Our idea consists in applying the χ^2 test to select AV cluster pairs relevant to the structure of the video from the N-best list of pairs obtained with mutual information. In other words, we want to find cutting points in the dendrograms where the resulting sets of audio and visual segments coincide more than at random. However, contrary to Dielmann's work, no labels are available and clustering does not provide a partition of the data into several classes. Therefore, applying the χ^2 test statistic as described in the previous section is not straightforward. Rather, we consider binary labels instead of an arbitrary number of labels obtained from a partitioning step. Let us consider the audiovisual segmentation resulting from the union of the audio and visual boundaries. Considering one AV cluster pair (i.e., a node from the audio dendrogram and one from the visual one), each audiovisual segment can be labeled with a binary label indicating whether the segment belongs to the AV cluster or not. More precisely, for a given AV cluster pair, we analyze the co-occurrences between audio and visual segments by the chi-squared distribution with 1 degree of freedom, i.e., O is now a 2×2 matrix representing the observations for two binary variables. The frequency of each pair of labels ($i \in \{0,1\}, j \in \{0,1\}$) is computed as the number of corresponding audiovisual segments[3]. For instance, for a given AV cluster pair, 0_{11} is the number of audiovisual segments belonging to both the audio and visual cluster, while 0_{10} indicates the frequencies of audiovisual segments that belong to the audio cluster but do not belong to the visual cluster. Since our objective is to determine whether there is a significant relationship between an audio cluster and a visual cluster, only the contribution of O_{11} to the test statistic χ^2 is verified. In particular, we compare a χ^2 distribution with a variable X_{11}^2:

$$X_{11}^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} \tag{4}$$

A threshold on the value of X_{11}^2 (set to 15 in our experiments) is used to decide if a given audiovisual cluster is relevant, i.e. jointly occur with sufficient frequency to be of interest for structure analysis. It can be observed that the χ^2 statistic is not reliable if the expected frequencies (E_{ij}) are too small (this problem was not mentioned in [11]). In our experiments, we reject pairs of labels (i.e., AV cluster pairs) for which $E_{11} < 1$. The χ^2 test is applied to all elements in the N-best list, yielding a filtered list of highly consistent pairs.

It is interesting to note that, at first glance, the mutual information (cf. Eq. 1) and the χ^2 test (cf. Eq. 4) encodes similar information and might therefore be redundant. However, Eq. 1 tends to select time-based consistent AV cluster pairs which generate time-based consistent segments, regardless of their frequencies. This may result in partially discovered events, i.e., events for which only a few occurrences were discovered, a fact that was experimentally confirmed. On the contrary, Eq. 4 selects AV cluster pairs based on the occurrence frequency only, which cannot ensure that all selected AV clusters are relevant. Therefore, both methods must be combined to ensure that the filtered N-best list contains AV cluster pairs consistent both in time-wise and frequency-wise.

[3]For practical reasons, very short segments of less than 10 ms are not counted.

6. EVENT CHARACTERIZATION AND MODELING

Because of the nested cluster structure of the dendrograms, the filtered list of AV cluster pairs contains redundant entries which correspond to the same underlying event. Typically, considering the parent (or the descendant) of either the audio cluster or the video cluster of a good AV cluster pair will most probably result also in a good AV cluster pair as the two (audio or visual) clusters, which differ only by one segment. It is therefore necessary to group redundant entries to define events for which models can be built in an unsupervised fashion.

Advantage is taken of the dendrogram structure to group AV cluster pairs which share segments. All AV cluster pairs belonging to the same branches in the two dendrograms are grouped together. The result is a list of non-intersecting groups, each of which represents a potential event.

Each potential event is thus characterized as a group of AV cluster pairs from which a model can be built by automatically selecting positive and negative examples. Let $E = \{e_1, e_2, ..., e_m\}$ be a group of AV cluster pairs, where e_i represents an AV cluster pair (C^A, C^V) with the corresponding temporal segments (S^A, S^V). Positive and negative samples for each pair $e_i \in E$ are determined as follows:

$$\vec{AV}_{s_k} \in +\mathbb{1} \text{ if } s_k \subset S^A \cap S^V$$
$$\vec{AV}_{s_k} \in -\mathbb{1} \text{ if } s_k \not\subset S^A \cup S^V$$

where s_k is an audiovisual segment as defined previously and $\vec{AV}_{s_k}$ is the corresponding audiovisual feature vector (i.e., the concatenation of the audio and visual features used for clustering). To select training samples for the group E, each element in the group cast its votes for negative and positive, and the accumulated results are kept for all elements. Thresholding is applied on the accumulated results of the votes to select positive and negative samples for the group. In our experiments, a (positive or negative) sample is selected if its votes are greater than the mean value of the corresponding accumulated voting result. In other words, we select as positive samples of an event those audiovisual segments which appear in most of the AV cluster pairs e_i. Similarly, negative samples correspond to audiovisual segments which appear rarely (if at all) in the AV cluster pairs characterizing the event. From the selected positive and negative examples, a binary SVM classifier is trained, yielding a model of the event which is classically used to detect the event considered in the video.

Fig. 2: Example of several typical structural events from the Canal9 data set. From left to right: a full group of participants, anchor person, and multiple participants.

order	"Baseline"			Our method		
	R	P	F1	R	P	F1
1st	0.95	0.59	0.69	0.94	0.77	**0.84**
2nd	0.94	0.63	0.73	0.87	0.75	**0.78**
3rd	0.92	0.58	0.66	0.81	0.77	**0.77**
4th	0.79	0.64	0.63	0.75	0.77	**0.71**
5th	0.74	0.60	0.57	0.60	0.73	**0.60**

Table 1: Multiple structural event detection performances.

7. EXPERIMENTAL RESULTS

Experiments are carried out on the standard, publicly available, Canal9 political debate data set (cf. Fig. 2), provided by Vinciarelli et al. [14] in 2009. To the best of our knowledge, this is the only publicly available data set that can be used to test audiovisual structuring tasks. This data set contains a collection of 72 political debates with roughly 42 hours of edited high quality audiovisual recordings, recorded by the Canal 9 local TV station and broadcast in Valais, Switzerland. Debates exhibit a strong audiovisual structure, with a limited number of speakers and a limited number of camera viewpoints. As a result, multiple audiovisually consistent events can be found in such videos. Typically such events are a close-up of one of the guests speaking or a global view of the guests with the anchor speaking.

Results are reported in terms of recall (R), precision (P), and F-measure (F1), computed on a time basis. An event discovered is first mapped to the corresponding reference event by finding the most overlapping event in the reference annotation. Given this mapping, recall is measured as the amount of time the discovery is correct divided by the total duration of the reference event. Precision is defined in a similar way. Recall and precision measures are averaged across files.

The output of the algorithm is a list of events discovered, ordered from the one with the best AV consistency according to the χ^2 test to the less. Results for the five best events discovered are reported in Table 1. For comparison purposes, results obtained without the χ^2 test (i.e., ranking of the events is solely based on the mutual information) are also reported (column "Baseline" in this table). Note however that, when not using the χ^2 test, we observed quite frequently candidate events for which SVM training could not converge. This is the case of partially discovered events, i.e., only a few occurrences have been detected, which lead to very few positive training samples selected in comparison with the negative ones. Results for such events are not included in the baseline performance which is therefore optimistic. From this table, one can see that the proposed method gives rather balanced recall and precision values while for the baseline, recall values are rather high whereas precision values are quite low. This reveals that each detected AV cluster pair from the "Baseline" experiments is not well matched to any annotated AV cluster, thus demonstrating the benefit of χ^2 filtering to select multiple events.

Besides this evaluation, we investigated how well typical events of such shows are discovered. To this end, we selected typical audiovisual events in the Canal9 data set, for example close ups of a guest speaking (see middle image in Figure 2). For such events, we search in the first 10 events discovered for a match and evaluate the match using recall and precision. Results are reported in Table 2 for different types of typical events, namely: Guest views which are individual participants where the person shown is also speaking; Full-group views which are specified by a whole group that appear when only one participant is speaking; Multiple participants consist of two or more participants which appear in different camera angles when one participant is speaking; Credits are like jingles in news videos, which appear at the beginning and the end of each debate; Topic introduction appear at the beginning of each debate, which comprise a generated-computer screen with background music. The last column in this table shows the most frequent rank at which the event was found in the list of events discovered. It can be seen from this table that our method achieves very good results on credits and topic introduction which exhibit very limited variability. Full-group views are poorly detected, primarily due to the high variations between audio clusters, i.e., for events of this genre, while the whole group is shown participants speaking in turns, resulting in strong variations between the occurrences. Finally, the last row of Table 2 reports results for compound event types where a group of participants is shown while several persons are speaking simultaneously. In this case, very poor results are obtained as grouping speech from multiple speakers is a highly challenging task. This illustrates the limits of our method for such challenging events.

Finally, we focus on events of type guest views, i.e., showing a guest speaking. Such events are of most importance in debates and accounts for 2/3 of the events annotated in the reference. Given a list of 10 events discovered, we evaluate for each rank the guest view events discovered for the rank. Table 3 reports recall and precision for each rank, as well as the cumulated ratio of guest view events discovered up to that rank. More than 70 % of the 50 guest view events of the reference are found before rank 6, with a precision of 76 % and

Event type	R	P	F1	Rank
Guest view	0.70	0.76	0.66	1
Multiple participants	0.53	0.71	0.53	8
Full-group views	0.46	0.72	0.50	7
Credits (jingles)	0.93	0.87	0.88	4
Topic introduction	0.78	0.99	0.86	6
Compound	0.64	0.16	0.21	7

Table 2: Average performances of the different genres of events for the first 10 events discovered.

Rank	R	P	F1	ratio
1st	0.94	0.77	0.84	34%
2nd	0.90	0.77	0.81	48%
3rd	0.87	0.77	0.80	56%
4th	0.85	0.77	0.78	68%
5th	0.80	0.76	0.75	74%
6th	0.77	0.76	0.72	76%
7th	0.74	0.76	0.70	76%
8th	0.72	0.76	0.69	76%
9th	0.70	0.76	0.66	80%

Table 3: Performance for the detection of Guest views events in the first 10 events discovered.

a recall of 80 %. Precision decreases quite moderately at each rank. These results demonstrate the benefit of unsupervised mining of multiple audiovisually consistent events for video structure analysis.

8. CONCLUSION

In this paper, we have presented a new framework that incorporates Pearson's χ^2 statistical test into a cluster selection method based on mutual information for the discovery of multiple audiovisual events in a video. Through experiments, we have shown that it is feasible to detect multiple events in a totally unsupervised way, without any prior knowledge on the events to be detected. We have demonstrated that the method can be used for video structure analysis. Future work includes the estimation of the number of events of interest and the use of more complex features for the discovery of events in different domains.

9. REFERENCES

[1] C.G. M. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools Appl.*, vol. 25, pp. 5–35, January 2005.

[2] T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. HSu, "Story boundary detection in large broadcast news video archives: techniques, experience and trends," in *Proceedings of the 12th ACM Multimedia*, pp. 656–659.

[3] F. Wang, Y.-F. Ma, H.-J. Zhang, and J.-T. Li, "A generic framework for semantic sports video analysis using dynamic bayesian networks," *Multi-Media Modeling Conference, International*, vol. 0, pp. 115–122, 2005.

[4] Y. Li, Shrikanth S. Narayanan, and C.-C. Jay Kuo, "Adaptive speaker identification with audiovisual cues for movie content analysis," *Pattern Recogn. Lett.*, vol. 25, pp. 777–791, May 2004.

[5] M. Covell, S. Baluja, and M. Fink, "Detecting ads in video streams using acoustic and visual cues," *Computer*, vol. 39, pp. 135–137, December 2006.

[6] C. Herley, "Argos: automatically extracting repeating objects from multimedia streams.," *IEEE Transactions on Multimedia*, vol. 8, no. 1, pp. 115–129, 2006.

[7] X.-F. Yang, Q. Tian, and P. Xue, "Efficient short video repeat identification with application to news video structure analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 3, pp. 600–609, 2007.

[8] A. Divakaran, K.A. Peker, R. Radhakrishnan, Z.Y. Xiong, and R. Cabasson, "Video summarization using mpeg-7 motion activity and audio descriptors," in *Video-Mining*. 2003, p. Chapter 4, Springer.

[9] Y. Wang, H. Jiang, M. S. Drew, Z. Li, and G. Mori, "Unsupervised discovery of action classes," in *Proceedings of the 2006 IEEE CVPR*, 2006, pp. 1654–1661.

[10] C. Ma and C. Lee, "Unsupervised anchor shot detection using multi-modal spectral clustering," in *ICASSP*, 2008, pp. 813–816.

[11] A. Dielmann, "Unsupervised detection of multimodal clusters in edited recordings," in *IEEE International Workshop on Multimedia Signal Processing (MMSP'2010)*, Saint-Malo, France, October 4-6 2010.

[12] M. Ben and G. Gravier, "Unsupervised mining of audiovisually consistent segments in videos with application to structure analysis," in *IEEE ICME'11*, Barcelona, Spain, July 2011.

[13] A.-P. Ta, M. Ben, and G. Gravier, "Improving cluster selection and event modeling in unsupervised mining for automatic audiovisual video structuring," in *The 18th Int. Conf. on MultiMedia Modeling*, Klagenfurt, Australia, 2012.

[14] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A database of political debates for analysis of social interactions," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2009.