A Multi-stage Approach for Anchor Shot Detection

L. D'Anna¹, G. Marrazzo¹, G. Percannella¹, C. Sansone², and M. Vento¹

¹ Dip. di Ingegneria dell'Informazione ed Ingegneria Elettrica, Università degli Studi di Salerno
Via Ponte Don Melillo, I, I-84084, Fisciano (SA), Italy
{ldanna, gmarrazzo, pergen, mvento}@unisa.it

² Dipartimento di Informatica e Sistemistica, Università degli Studi di Napoli "Federico II"
Via Claudio 21, I-80125 Napoli, Italy
carlo.sansone@unina.it

Abstract. In this paper we present a novel algorithm for anchor shot detection (ASD). ASD is a fundamental step for segmenting news video into stories that is among key issues for achieving efficient treatment of news-based digital libraries.

The proposed algorithm creates a set of audio/video templates of anchorperson shots in an unsupervised way, then classifies shots by comparing them to the templates. Audio similarity is evaluated by means of a new index and helps to achieve better performance than a pure video approach. The method has been tested on a wide database and compared with other state-of-the-art algorithms, demonstrating its effectiveness with respect to them.

1 Introduction

Story segmentation is a basic step towards effective news video indexing. All the solutions to this problem proposed in the literature may be ascribed to one of the two following approaches. According to the first, segmentation is accomplished by directly finding the story boundaries. Such boundaries are typically obtained by looking for the occurrences of some specific event (a sequence of black frames, the co-occurrence of a silence in the audio track and a shot boundary in the video track, etc.), or an abrupt change of some features at a high semantic level, as a topic switch. The main limitation of this approach relies on the fact that the overall performance depends in the first case to the validity of the hypothesis that a story boundary is associated to a specific event in the audio or the video stream, while in the second case to the possibility of reliably deriving high semantic level features.

The other approach performs story segmentation according to the following news program model assumption: given that each shot of the news video can be classified as an anchor shot or a news report shot, then a story is obtained by linking each anchor shot with all successive shots until another anchor shot, or the end of the news video, occurs. Using this model for the stories, news boundaries correspond to a transition from a news report shot to an anchor shot, or from an anchor shot to another. According to the above news story model, automatic anchor shot detection (ASD) becomes the most challenging problem to partition a news video into stories. It has to be noted that the main limitation of this approach relies on the validity of the above described news story model. However, some papers [1,2] have shown that such

a model is valid for most TV networks. Consequently, we preferred to follow this approach, directing our efforts to provide a solution to the ASD problem.

In the scientific literature there are many papers that propose ASD algorithms: the majority exploits only video information. They can be roughly grouped into two categories. First approaches rely on the definition of a set of models of anchor shots, so that ASD is done by matching news video shots with the models [3,4,5]. Typically, a distinctive frame, called *key-frame*, is extracted for each shot and used for detection purposes. In [3] color classification and template matching are used. In [4] a unique anchor shot model is defined, while in [5] an anchor shot is modeled as a sequence of frame models. All these approaches strongly depend on the specific video program model. This is a severe limitation, since it is difficult to construct a general model able to represent all the different kind of news and since the style of a particular news program can change over the time.

In order to overcome this limitation, some authors [6] proposed to build a key-frame model in an unsupervised way. However, since a single model is chosen for each news video, the authors implicitly assume that different anchorperson models share the same background. This is not true for most news stations: because of different camera angles, different models can have different backgrounds. Other authors [2,7] propose unsupervised methods that look for shots with similar visual contents that repeatedly occur during the whole news video. In particular, in [2] a graph-theoretical cluster analysis method is employed to classify video shots. As pointed out by the authors, this approach fails when identical news report shots appear in different stories of the same news program, or when an anchor shot model is present once in a program (for example, the case in which the anchorperson appears for most times in the right or left part of the screen and only once at the center). In [7] shot classification is also performed on the basis of the motion features of the anchor shot. For the authors, it is reasonable to assume that in an anchorperson shot both the camera and the anchorperson are almost motionless. In some TV news, however, zooming effects can be used also in an anchor shot; if motion is contained also in the background, the anchor shot can be missed.

In the last years the use of audio as a good additional source of information for video segmentation has been rapidly raised up. There is, in fact, a number of systems that integrate audio and video features in the context of news segmentation by performing multi-modal analysis [8]. However, the majority of the presented proposals use audio features for directly individuating news boundaries, by means of a silence or a speaker change detector, in order to strengthen or to weaken the boundaries provided by the analysis based on video techniques.

Among the approaches that use the audio track for ASD, in [9] the authors propose a Hidden Markov Model (HMM) to classify frames on the basis of the statistics of the frames present in a news video. The features used are the difference image between frames, the average frame color and also the audio signal. The HMM parameters are evaluated during a training phase by using the ground truth of a given news program, and then are dependent on the style of the specific news video edition. In [10] it is proposed a technique that performs segmentation and clustering of portions of video with similar audio and video contents and tries to find temporal synchronization between pairs of clusters. In particular, ASD is performed by separately clustering and then comparing audio clips and video key-frames. When sufficient overlap is found between an audio and a video cluster then an anchor shot is detected. Unfortunately,

parallel audio and video clustering often lead to dissimilar grouping solutions, e. g. when a news report is commented by the anchorperson, or when there is a speaker change within a shot.

All summarizing, the major drawbacks of former approaches are the following: *i)* supervised model-based techniques are not general enough, as they require *a priori* definition and construction of an anchor shot model; *ii)* the definition of a unique anchor shot model for a news edition is restrictive and gives rise to missed detections when different backgrounds are present in a news video edition; finally, *iii)* indiscriminate use of audio information is not effective due to its incoherence with video and then yields to a misleading shot classification.

In this paper, we propose a two stage audio/video ASD method that is able to overcome all the above limitations. In the first stage the method builds in an unsupervised way a set of templates, each one representing a different anchor shot model within a video. The second stage uses a video similarity metric to retrieve a set of candidate anchor shots, which might have been missed by the first stage, and classify them by evaluating the audio similarity with respect to the templates. Note that, differently from the formerly described approaches, we do not use audio information as-it-is, but we perform audio-based classification only on the set of candidate shots and by employing a suitably defined similarity metric.

We tested our method on a significant database made up of several video news editions from the two main Italian broadcasters (RAI 1 and CANALE 5), obtaining a very good performance. Moreover, we also compared our algorithm with other three state-of-the-art unsupervised ASD algorithms achieving significant performance improvements.

The organization of the paper is as follows: in Section 2, the proposed algorithm is described; in Section 3, the database used is reported together with the tests carried out in order to assess the performance of the proposed algorithm; finally, in Section 4, some conclusions are drawn.

2 The Proposed Approach

As stated in the introduction, we propose a two stage analysis which is able to achieve ASD in an unsupervised way. The first stage extracts a set of audio/video templates from the news video under analysis, so avoiding any training procedure or manual definition of the templates. The second stage selects a set of candidate anchor shots by means of a video similarity metric with respect to the templates, then validates them by exploiting both audio similarity and the presence of faces. Details about template nature and similarity metrics used for shot classification will be given in next subsections.

2.1 First Stage: Anchor Shot Template Extraction

The preliminary task is to define and build a set of templates as audio/video touchstones for all the shots in the TV news program. The main difference with the previous papers in the literature is the definition of several anchor shot templates: in fact, during a news program there are typically several kinds of anchor shot camera settings, due to different angle, background and distance with respect to anchor

person. A single anchor shot template would not match all these situations and would introduce many missed items in the anchor shot detection task. In our approach each template practically corresponds, from the video point of view, to a single frame of the considered shot (i.e., its key-frame), while, from the audio point of view, to the audio shot characterization in a given feature space.

After shot segmentation, all video shots are processed in order to build the templates. As anchor shots are mainly identifiable from their high visual similarity, we preliminarily need a clustering technique in order to group similar shots and find candidate anchor shot clusters. Then, several heuristics can be used to discard false detected clusters: anchor shots occur at least two times with the same angle, have a large temporal spanning along the news program and are characterized by the presence of a face. All these features are taken into account in the extraction of the anchor shot templates.

A first clustering task groups together shots with same visual appearance. We used a graph-theoretical clustering (GTC) analysis, which considers shot key-frames as nodes of a complete graph in a given feature space. Each edge of the graph is assigned a weight corresponding to a distance between pairs of nodes, then the *minimum spanning tree* (MST) is built on the graph. The distance between nodes is defined in this way: each key-frame is divided into 16 rectangular regions of the same size, then color histograms of corresponding regions of the two nodes are compared; the eight regions with the most similar histograms are individuated and finally the sum of their histograms differences is considered as the distance between nodes. After constructing the MST and removing all the edges in the tree with weights greater than a threshold λ , a *forest* containing a certain number of subtrees (*clusters*) is obtained. Each cluster correspond to a group of visually homogeneous shots. As anchor shots occur repeatedly during a news edition, clusters with less than 2 nodes are discarded. For determining the optimal value of λ , we used a *Fuzzy C-Means* clustering algorithm [11]. This algorithm builds a list of all edges, sorted by their weights, and finds the best separation threshold which partitions the list in two parts. Edges belonging to the sub-list with the highest weights are pruned.

The second step discards the clusters with low *lifetime*, i.e. the time interval that includes all the shots of the cluster. Anchor shot occurrences are typically temporally sparse, so clusters with lifetime lower than a threshold δ are removed. Threshold setting can be performed in a straightforward way, on the basis of the length of the specific news program: some details about this point are provided in the next section. In addition, it is worth noting that this setting is not critical, since anchor shots missed because of a non optimal setting of the threshold, can be recovered by the successive stage. Lifetime analysis is very effective in managing interviews during news reports, where recurrent camera shots on the interviewed person may generate large clusters; checking the lifetime allows to classify the shots within such clusters as news reports since their lifetime is typically very small.

The last step discards the shots without faces. We introduced a robust face detection method which requires presence of a face all along a shot. To this aim we extract three frames from each shot (the first, the middle and the last frames) and apply the face detection algorithm [12] on them. If there are more than one frame without faces then the shot is removed.

After this step, we can again eliminate clusters with less than 2 shots, since they violate the assumption that an anchorperson occurs at least twice in the video. Moreover, some clusters may have changed their lifetime, so we apply again the lifetime control. This procedure gives rise to the final set of anchor shot clusters.

Finally, for each remained cluster we build a set N_{As} containing all the shots belonging to that cluster and extract a unique key-frame from each N_{As} : these key frames represent our anchor shot templates from a video point of view. Moreover, from an audio point of view, we assume that the visual presence of the anchorperson corresponds to his/her speech. Consequently, we assume as audio template the union of the audio portions relating to all the shots belonging to the set N_{As} , represented in an adequate feature space. Figure 1 summarizes the complete template extraction algorithm.

- 1) Construct a complete graph G such that:
 - a) its nodes Kf_i correspond to the key-frames of each shot;
 - b) each of its edges $e_{ij}=(Kf_i, Kf_j)$ is characterized by a weight $w_{ij}=d(Kf_i, Kf_j)$, where $d(Kf_i, Kf_j)$ is calculated on the basis of the color histogram differences between Kf_i and Kf_j ;
- 2) determine the minimum spanning tree (MST) of G ;
- 3) remove from the MST those edges with large weights by using the *Fuzzy C-Means* algorithm, in order to create shot clusters;
- 4) remove clusters with only one node;
- 5) remove all clusters with lifetime lower than a threshold δ ;
- 6) extract 3 frames from each shot S_j belonging to the set of remaining clusters;
- 7) remove from clusters those shots which have less than 2 frames which contain a face;
- 8) apply steps 4) and 5) to remaining clusters;
- 9) for each remained cluster build the set N_{As} by extracting all anchor shots from that cluster.
- 10) extract a unique key-frame from each cluster and the whole audio track from the set N_{As} , giving rise to the set of audio/video anchor shot templates.

Fig. 1. The proposed template extraction algorithm

2.2 Second Stage: Shot Classification

Classification must be performed on all the shots outside the final set of clusters, as the first stage might have missed those anchor shots occurring only once along the news program or which have low lifetime. These shots can be seen as missed templates of anchor shots which correspond to special camera settings, due to lighting, angle or special background with respect to the anchorperson. However, we observed that during a news edition there are only few of such special kinds of anchor shot, so we expect that only few candidate shots need to be recovered. Consequently, we perform a preselection of these candidates on the basis of the video similarity with respect to the already found templates. Finally, the candidates are classified by both audio and face detection. This framework prevents us from performing audio classification on the whole set of shots, as audio processing is highly time-consuming. Furthermore, audio classification would also bring to false detected anchor shots, due to cases where the anchorperson directly comments a news report.

During the candidate selection step, we consider one template at a time and compare it to the key-frame of each of the discarded shots: then, we select only the three

candidate shots with the highest similarity with respect to any of the templates. To define similarity we use the metric presented in [13], which is computationally efficient and sensible to global features such as the general studio setting. The need for a more global metric lies in the fact that if we used the same technique as in the clustering step, we would discard the same shots as in the first phase; moreover, a more global similarity metric is more permissive and let us take into account, for the classification task, also those shots which are not so strictly related to the templates.

Audio classification requires shot characterization in an adequate feature space. By using the results reported in [14], a set of 48 features (20 MFCC, 14 LPCC, 14 PF, see [14] for further details) is extracted from each frame in the audio track of the shots. In this case, a frame is a segment of 1024 audio samples.

After feature extraction, each shot is filtered in order to remove non-voiced and silence portions by means of an appropriate masking module. This is a useful step because voiced segments correspond to vocal cords movement, so they characterize a speaker from an acoustic point of view. Moreover, filtering out audio portions helps in speeding-up the shot classification process. Our masking module is based on the analysis of Energy and ZCR (Zero Crossing Rate) and segments audio into three classes of clips: voiced, non-voiced and pause. Our method is an improvement of the method presented in [15], obtained by discarding frequent audio class transitions and by merging those voiced clips which are separated by a single pause segment.

Audio shot classification is carried out by computing for each shot to be classified the value of an adequate similarity index, namely *D-index*, which expresses similarity between a shot and a template from an audio point of view.

Each shot S_i is considered as a cluster of audio feature vectors, so it can be represented by its centroid, namely C_i . For a generic pair of shots (S_m, S_k) belonging to N_{As} we assume:

$$D_{m,k} = 1 - \frac{d(C_m, C_k)}{d_{\max}} \quad (1)$$

where $d(C_m, C_k)$ is the Euclidean distance between C_m and C_k , while $d_{\max}$ is the diameter of the cluster of centroids of the shots belonging to N_{As} .

Candidate selection

- 1) Select a template;
- 2) compare all the discarded shots with the template;
- 3) build the candidates list L_c of the three most similar shots (according to a suitably defined similarity metric), sorted in descending order;
- 4) repeat steps 1), 2) and 3) for every template and refresh L_c ;

Classification

- 5) Extract audio feature vectors for all shots in N_{As} and L_c ;
- 6) remove unvoiced and silent segments by means of a masking module;
- 7) calculate D_i for each candidate shot S_i , by comparison with the audio templates and make audio classification decision for S_i ;
- 8) apply face detection to candidates;
- 9) apply AND rule between face detection and audio classification to reach the final decision.

Fig. 2. The shot classification algorithm

Consequently, we can consider d_{max} as the diameter of the set of templates. Given a generic shot S_i to be classified, we calculate its *D-index*, namely D_i , as the average of all the $D_{i,k}$ obtained by considering all the shots S_k in the set N_{As} . If $D_i > 0$ then S_i is classified as anchor shot. The case $D_i > 0$ implies that the average distance between the shot S_i and the set of templates is lower than the diameter d_{max} of the set of templates. Consequently, we classify S_i as an anchor shot. On the contrary, if $D_i < 0$ then the shot S_i is sufficiently far away from the cluster of templates, so it is discarded as a news report shot.

Finally, a further face detection module helps to validate the classification, so that only those candidates which are classified as anchor shots by both audio similarity and face detection are stated as anchor shots.

A scheme of the classification stage is provided in Figure 2.

3 Experimental Results

In order to assess the performance of our approach, we have collected a database composed by several videos from the two main Italian broadcasters, namely, RAI 1 and CANALE 5. Our database includes the main news editions from each broadcaster. A similar dataset, developed by the Linguistic Data Consortium for the TREC Video Retrieval Evaluation contest and composed of 70 hours of news video from ABC and CNN, was unfortunately not publicly available. Table 1 shows some details of the considered database.

Table 1. The video database used in this paper

Broadcaster	No. editions	Total length	No. anchor shots	No. news reports
RAI 1	26	7:32:19	377	4602
CANALE 5	16	9:21:12	250	4269
TOTAL	42	16:53:31	627	8871

The performance of our system is expressed in terms of *Precision* and *Recall* [16]. The *F-measure* has also been used, since it combines the former indexes in a single figure of merit according to the following formula:

$$F = (2 * Precision * Recall) / (Precision + Recall).$$

In Table 2 the performance of the first stage alone, considered as a preliminary classification stage, are reported.

Table 2. Performance after the first stage of the proposed method

	<i>Precision</i>	<i>Recall</i>	<i>F</i>
RAI 1	0.994	0.924	0.957
CANALE 5	0.967	0.891	0.927

The second stage classifies the set of three candidates, so recovery is effective only if missed (or “recoverable”) anchor shots are selected as candidates. The performance of the second stage is shown in Table 3. In order to evaluate the effectiveness of the audio similarity method proposed in this paper, in Table 3 we have reported also the highest achievable performance, calculated as that obtained if the second stage introduced no further false item and recovered all missed anchor shots. Note that almost all anchor shot are recovered, while only few new falsely detected anchor shots were introduced.

Table 3. Performance of the second stage of the proposed method

	<i>Recovered / Recoverable anchor shots</i>	<i>Falses / Introducible falses</i>	<i>Precision / Max Precision</i>	<i>Recall / Max Recall</i>	<i>F / Max F</i>
<i>RAI 1</i>	19/19	5/29	0.968/0.994	0.979/0.979	0.974/0.987
<i>CANALE 5</i>	14/15	7/63	0.940/0.969	0.952/0.956	0.946/0.963

During experimental phase we had to tune the threshold δ on the lifetime value. We verified that δ can change in a wide range with almost no effect on performance, and that its optimal value depends only on the length of the news edition. We fixed $\delta=2'$ for news editions shorter than fifteen minutes and $\delta=4'$ otherwise. In this sense, our algorithm can be really considered as unsupervised.

The proposed method has also been compared with three state-of-the-art unsupervised ASD algorithms [2, 6, 7]. Each of these algorithms is characterized by several thresholds, so different operating points can be obtained in a *Precision-Recall* plane [16]. We decided to choose the values of the thresholds that maximize F over the whole set of videos. This has been done separately for each of the two TV-networks. Obviously, this is an overestimation of the real performance of the algorithms, since such maximization should be done on a different set of news videos. It is also worth noting that, as experimentally demonstrated in [17], the choice of the operating point is crucial for the algorithms, as their performance dramatically depends on the choice of the thresholds.

Table 4. Performance comparison of our algorithm with respect to the algorithms in [2, 6, 7]

	RAI 1			CANALE 5		
	<i>Precision</i>	<i>Recall</i>	<i>F</i>	<i>Precision</i>	<i>Recall</i>	<i>F</i>
Our algorithm	0.968	0.979	0.974	0.940	0.952	0.946
Gao and Tang [2]	0.827	0.928	0.875	0.810	0.889	0.848
Hanjalic et al. [6]	0.681	0.617	0.647	0.889	0.533	0.667
Bertini et al. [7]	0.987	0.822	0.897	0.867	0.796	0.830

In Table 4 the performance of our algorithm is compared with those obtained by the methods in [2, 6, 7] on our dataset, so demonstrating its effectiveness with respect to them.

4 Conclusions

In this paper a novel algorithm for anchor shot detection is presented. An effective audio/video anchor shot template matching algorithm is introduced, in order to gain effectiveness against unavoidably missed items left out by pure video analysis. Moreover, a new audio similarity index is discussed, which allows the definition of a synthetic and unique value to quantify resemblance of a generic shot to the template from an audio point of view. The method has been tested on a news video database consisting of about 20 hours, providing significant improvements with respect to other state-of-the-art algorithms.

Future work will include a refinement in the anchor shot audio recognition module, which will be able to detect whether there is one or two anchor persons, so to consequently define a single or a pair of values for the similarity index for each shot.

References

1. M. De Santo, G. Percannella, C. Sansone, M. Vento, "An Unsupervised Shot Classification System for News Video Story Detection", in A.F Abate, M. Nappi, M. Sebillo (eds.) *Multimedia Database and Image Communication*, World Scientific Publ., pp. 93-104, 2005.
2. X. Gao, X. Tang, "Unsupervised Video-Shot Segmentation and Model-Free Anchorperson Detection for News Video Story Parsing", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, No. 9, pp. 765-776, 2002.
3. B. Günsel, A. M. Ferman, A. M. Tekalp, "Video Indexing Through Integration of Syntactic and Semantic Features", *Proc. of Workshop Applications of Computer Vision*, Sarasota, FL, pp. 90-95, 1996.
4. D. Swanberg, C.F. Shu, R. Jain, "Knowledge Guided Parsing in Video Databases", *Proc. of SPIE Symposium on Electronic Imaging: Science and Technology*, San Jose, CA, pp. 13-24, 1993.
5. S. W. Smoliar, H. J. Zhang, S. Y. Tao, Y. Gong, "Automatic Parsing and Indexing of News Video", *Multimedia Systems*, vol. 2, no. 6, pp. 256-265, 1995.
6. A. Hanjalic, R. L. Lagendijk, J. Biemond, "Semi-Automatic News Analysis, Indexing, and Classification System Based on Topics Preselection", *Proc. of SPIE, Electronic Imaging: Storage and Retrieval of Image and Video Databases*, San Jose (CA), 1999.
7. M. Bertini, A. Del Bimbo, P. Pala, "Content-Based Indexing and Retrieval of TV News", *Pattern Recognition Letters*, vol. 22, pp. 503-516, 2001.
8. C.G.M. Snoek, M. Worring, "Multimodal Video Indexing: A Review of the State-of-the-art", *Multimedia Tools and Applications*, vol. 25, pp. 5-35, 2005.
9. S. Eickeler, S. Muller, "Content-based video indexing of TV broadcast news using Hidden Markov Models", *ICASSP '99*, pp. 2997-3000, 1999.
10. W. Qi, L. Gu, H. Jiang, X. R. Chen, H. J. Zhang, "Integrating Visual, Audio and Text Analysis for News Video", *7th IEEE International Conference on Image Processing*, Vancouver, British Columbia, Canada, 2000.
11. J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York, 1981.
12. P. Viola, M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features", *Proc. of the IEEE CVPR Conference*, vol. 1, pp. 511-518, 2001.
13. H. Y. Lee, H. K. Lee, Y. H. Ha, "Spatial Color Descriptor for Image Retrieval and Video Segmentation", *IEEE Transactions on Multimedia*, vol. 5, no. 3, pp. 358-367, 2003.

14. L.P. Cordella, P. Foggia, C. Sansone, M. Vento. "A Real-Time Text-Independent Speaker Identification System". 12th International Conference on Image Analysis and Processing, IEEE Computer Society Press, Mantova, Italy, pp. 632 - 637, 17 - 19 September, 2003.
15. D. Wang, L. Lu, H-J Zhang, "Speech Segmentation Without Speech Recognition", ICASSP 2003 vol. I, pp.468-471, 2003.
16. U. Gargi, R. Kasturi, S.H. Strayer, "Performance Characterization of Video-Shot-Change Detection Methods", IEEE Trans. on Circuits and Systems for Video Technology, vol. 10, no. 1, pp. 1-13, 2000.
17. M. De Santo, G. Percannella, C. Sansone, M. Vento, "A Comparison of Unsupervised Shot Classification Algorithms for News Video Segmentation", Lecture Notes in Computer Science vol. 3138, Springer, Berlin, pp. 233-241, 2004.