An Improved Algorithm for Anchor Shot Detection

M. De Santo¹, G. Percannella¹, C. Sansone², and M. Vento¹

¹ Dipartimento di Ingegneria dell'Informazione e di Ingegneria Elettrica,
Università di Salerno - Via P.te Don Melillo, 1 I-84084, Fisciano (SA), Italy
{desanto, pergen, mvento}@unisa.it

² Dipartimento di Informatica e Sistemistica, Università di Napoli "Federico II",
Via Claudio, 21 I-80125 Napoli (Italy)
carlosan@unina.it

Abstract. Segmentation of news videos into stories is among key issues for achieving efficient treatment of news-based digital libraries. Indeed, anchor shot detection is a fundamental step for segmenting news into stories.

In this paper we present an improved algorithm for anchor shot detection. It is based on a graph theoretical clustering method and exploits the idea of the *cluster lifetime* for improving the performance. Moreover, a method for automatically fixing all the thresholds required by the original version of the algorithm is also proposed, so making it fully unsupervised.

The proposed algorithm has been tested on a database significantly wider than those typically used in the field, demonstrating its advantages with respect to the original version.

1 Introduction

Among all the different sources of video material nowadays available, news videos received great attention by the scientific community. This is mainly due to the fact that broadcasters are interested in building large digital databases of their resources, so as to allow the reuse of the archived material for other TV programs. Such a reuse can be efficiently performed only after a suitable indexing procedure of the existing materials, where an important step towards effective indexing is the segmentation of a news video into *stories*.

At a first stage, this implies the partition of the video into *shots*, i.e. sequences of frames, obtained by detecting transitions that are typically associated to camera changes. Once shots have been individuated, they can be classified on the basis of their content. Two different classes are typically considered: *anchor* shots and *news-report* shots. Successively, the given news video can be segmented into stories. This is obtained by linking each anchor shot with all successive news report shots until another anchor shot, or the end of the news video, occurs. Using this model for the stories, anchor shot detection algorithms can be employed as basic step for segmenting news videos.

In the literature, most of the approaches to the anchor shot detection problem exploit the video source information by using a model matching strategy [1]. For each shot, a distinctive frame, called *key-frame*, is extracted. Then, it is matched against a

set of predefined models of an anchor shot frame in order to classify it. This approach is strongly dependent on the model of the specific video program. This is a severe limitation, since it is difficult to construct all the possible models for the different news videos and the style of a particular news program can change over the time.

Other authors use a face detection approach to identify anchor shots [2]. However, face detection in video is generally too time-consuming for practical application. Furthermore, a shot where a reporter (not an anchorman!) is present can be erroneously recognized as an anchorperson shot by the face detection module.

In order to overcome the above reported limitations, some authors [3,4,5] propose methods that are unsupervised and do not require the explicit definition of an anchor shot model. Among them, one of the most effective algorithms is that proposed by Gao and Tang in [3]. Here, anchor shots are identified by using an algorithm based on graph-theoretical cluster analysis. It automatically groups similar key-frames into clusters, on the basis of their color histograms. The key-frames composing a cluster are classified as potential anchorperson frames if the cluster size is greater or equal to 2. Then, a spatial difference metric (SDM) is used to refine the shot classification; in fact, in some situations, the key-frames in a cluster may have similar histograms but different content. If a cluster has an average SDM value higher than a suitable threshold, it is removed from the anchorperson frame list. As pointed out by the authors, however, this approach fails when identical or very similar news-report shots appear in different stories of the same news program.

A more recent version of the algorithm has been proposed by Gao *et al.* in [6]. Here, the idea presented in [5] is adopted for reducing the number of false alarms (i.e., falsely detected anchor shots). In particular, starting from the key-frames of the cluster with maximal size, a template is generated and used to remove clusters having key-frames significantly different from it. The template is obtained in an unsupervised way, so preserving one of the most distinctive features of the algorithm. However, this approach cannot be used when there are two anchorpersons in the news video. In these cases, in fact, there are at least three different anchorperson models, while the template obtained by following the approach proposed in [5] is unique for each news video.

In this paper we propose an improved version of the algorithm presented in [6] that uses the idea of the *cluster lifetime* (borrowed from [4]) to deal with the false alarms provided by the algorithm proposed by Gao and Tang. In particular, the temporal interval (*lifetime*) that includes all the occurrences of key-frames belonging to a cluster is evaluated. Since anchorperson shots repeatedly occur along the whole video, the clusters having a lifetime smaller than a suitably fixed threshold are removed from the anchorperson list.

Moreover, as noted by the authors in [6], the algorithm proposed by Gao and Tang needs that some parameters are specified in advance. In this paper we also propose a method for automatically fixing the two thresholds required by the original version of the algorithm and experimentally prove that the threshold on the lifetime value can be fixed in a straightforward way.

In order to test the modified algorithm in a significant way, we built-up a database that is considerably bigger than those typically used in the field [3]. Namely, we used a news video database consisting of about 17 hours with 673 anchor shots and 8922 news report shots. Since in this database there are also news videos presented by two anchorpersons, in our tests we consider as benchmark only the first version of the algorithms proposed by Gao, i.e. the one presented in [3].

The organization of the paper is as follows: in section 2, the original version of the algorithm is recalled and the proposed modifications are presented. 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 Algorithm

The authors in [3] propose to classify video shots by using an algorithm based on graph-theoretical cluster (GTC) analysis. More in details, they propose an anchor shot detection scheme composed of four steps: short shot filtering, key-frame extraction, GTC analysis and post-processing.

In general, an anchorperson shot should last for more than 2 sec, since this shot should involve at least one sentence pronounced by the reporter. Therefore, if a shot lasts less than 2 sec it is considered as a news report shot. Otherwise, it is further analyzed through later steps.

The second step is the key-frame extraction: the authors propose that the middle frame is taken as the key-frame. These key-frames are the input to the GTC analysis module. It considers them as vertices in a feature space and then builds the minimum spanning tree (MST) [8] on these vertices. For constructing the MST it is necessary to associate a weight to each edge connecting two vertices. So, a distance between two key-frames needs to be defined. This distance is the weight associated to the edge that connects the two vertices representing the key-frames in the feature space. Each key-frame is divided into 16 regions of the same size; the histograms of corresponding regions in the two key-frames are compared and the eight regions with the largest histogram differences are discarded to reduce the effects of object motion and noise. The distance between these two key-frames is then defined as the sum of the histogram differences of the remaining regions. By using this distance, the MST can be constructed. Successively, by removing all the edges in the tree with weights greater than a threshold γ , a *forest* containing a certain number of subtrees (*clusters*) is obtained. In this way, the GTC method automatically groups similar vertices (key-frames) into clusters. The key-frames composing a cluster are classified as potential anchorperson frames if the size of the cluster is greater or equal to 2.

Starting from this set of potential anchorperson frames, the last step of the proposed detection scheme operates a further filtering. In fact, in some situations, the key-frames in a cluster may have similar histograms but different content. To detect this situation, a spatial difference metric (*SDM*) between two key-frames *KF1* and *KF2* in a cluster is proposed:

$$SDM = \frac{1}{M \times N} \sum_{i=1}^M \sum_{j=1}^N |I_{KF1}(i, j) - I_{KF2}(i, j)| \quad (1)$$

where $I_{KF1}(i, j)$ and $I_{KF2}(i, j)$ denote the intensity of a pixel at location (i, j) in the frames *KF1* and *KF2* respectively, and the frame size is $M \times N$.

If a cluster has an average *SDM* higher than a threshold λ , the whole cluster is removed from the anchorperson frame list. By using the *SDM* filtering, the final anchorperson shot classification is obtained.

It is worth noticing that the above anchorperson shot detection scheme requires to specify in advance two thresholds: γ for the GTC algorithm and λ for the post-processing step, respectively.

2.1 The Modified Version

As anticipated in the introduction, the main contribution of this paper is twofold: on one side we try to improve the performance of the algorithm proposed by Gao and Tang by removing some false alarms it produces, while on the other side we propose a method for automatically fixing the two thresholds γ and λ required by the original version of the algorithm, so making it completely unsupervised.

In order to remove the false alarms generated by the algorithm proposed by Gao and Tang, we add a further filter after the post-processing stage. This filter is based on the idea of the *cluster lifetime* (borrowed from [4]). It is defined as the temporal interval that includes all the occurrences of key-frames belonging to a cluster. The rationale of this kind of filter lies in the observation that anchorperson shots repeatedly occur along the whole video. Therefore, clusters having a lifetime smaller than a suitably fixed threshold δ should not belong to the anchor class and then can be removed from the anchorperson list.

This filter is very effecting in dealing with the very frequent situation of a person interviewed: in this case there are at least two shots (Fig. 1.a and 1.b) whose key-frames are very similar each other, separated by one or more shots in which there is the presence of a reporter (or the anchorman himself). The shots represented by similar key-frames are grouped together into a cluster by the GTC and the average SDM value of the cluster is very low: as a result the original algorithm erroneously attributes the shots to the anchor shot class. However, the proposed filter recognizes these shots as a news report since their cluster lifetime is very small, as it can be noted by considering the difference between the indexes of their key-frames shown in Fig. 1.

Fig. 1. Key-frames extracted from two news videos of our database that are erroneously detected as anchor shots by the original algorithm, but correctly classified by the modified version proposed here

It is worth noting that the filter based on the lifetime also uses a threshold; however, as already noted in [4], the threshold on the lifetime value can be fixed in a straightforward way on the basis of the length of the specific news program (more details about this point will be given in the experimental section).

The second contribution provided in this paper consists in a method for automatically computing the two thresholds γ and λ required by the original version of the algorithm.

As regards γ , our proposal is to determine its value by reformulating the problem as the one of partitioning the whole set of edges into two clusters, according to their weights. The cluster of the edges of the MST with small weights will contain edges to be preserved, while the edges belonging to the other cluster will be removed from the MST. In order to solve this problem we employ the Fuzzy C-Means (FCM) clustering algorithm.

FCM is a clustering technique based on the minimization of the following objective function:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m (x_i - c_j)^2, \quad 1 \leq m < \infty \quad (2)$$

where m is any real number greater than 1, x_i is the i -th measured data (in our case the weight of the i -th edge of the MST), c_j is the center of the cluster, u_{ij} is the degree of membership of x_i to the cluster j , C is the number of clusters (in our case $C = 2$) and N is the number of objects to be clustered. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{x_i - c_j}{x_i - c_k} \right)^{\frac{2}{m-1}}} \quad \text{and} \quad c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m} \quad (3)$$

This iteration will stop when:

$$\max_{ij} \left\{ |u_{ij}^{(k+1)} - u_{ij}^{(k)}| \right\} < \varepsilon \quad (4)$$

where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m . At the end of the procedure, each edge x_i has been assigned to the cluster r such that:

$$r = \arg \max_j u_{ij} \quad (5)$$

At this point, all the edges of the MST are separated into two clusters. Then, we remove from the MST all the edges belonging to the cluster s whose center exhibits the largest value, i.e.:

$$s = \arg \max_j c_j \quad (6)$$

As regards the threshold λ , we consider the K clusters obtained by using the GTC analysis in increasing order with respect to the value of the *SDM*. Let $SDM(0)$ be the smallest value and $SDM(K-1)$ the largest values. Then, λ is calculated as the average value of the first N *SDM* values, where the parameter N is simply related to the number of anchorpersons of each news program. In fact, we use $N = 4$ if the news video has only one anchorperson, otherwise $N = 6$. This choice is justified by the fact

that in the editions presented by two anchorpersons there are typically two additional anchor shot models with respect to the editions with a single anchorperson.

In order to strengthen the method against possible SDM outliers, we prefer to discard $SDM(0)$ during the computation of the average value. So, the value of λ is calculated as follows:

$$\lambda = \frac{1}{N-1} \sum_{i=1}^{N-1} SDM(i) \tag{7}$$

However, this procedure cannot be effectively applied when the number of clusters resulting from the GTC analysis is equal or even less than three. In a news video, in fact, there are in general at least three different anchor shots models. In this case we do not activate the SDM filtering; the cluster lifetime will discard false clusters, if any.

3 Experimental Results

Some efforts have been spent in the recent past by other researchers in building video databases for benchmarking purposes; in particular in [7] a database was built in order to characterize the performance of shot change detection algorithm. This database, however, is not adequate for our aims, since it is made up not only of news videos but also of sport events and sitcom videos, and the duration of news videos is only 20 minutes. To reproduce as much as possible the variability of the phenomenon under study, different news video editions of a single broadcaster should be considered, as well as news videos of different broadcasters. For this reason, the database used in this paper (about 17 hours) is composed by 28 news videos from the main Italian public network (namely, RAI 1) and 17 videos from the main Italian private network (namely, CANALE 5). Particular care was taken in order to include in the database the main news editions from these broadcasters. As it can be easily noted from Table 1, the size of our database is large; this is more evident if it is compared with the database used by Gao and Tang in the paper [3]. The performance of the proposed algorithm was then assessed on the two TV-networks, separately. Note that all the editions of RAI 1 are presented by a single anchorperson, while the news videos of CANALE 5 are presented by two anchorpersons, even if some anchor shots of these news videos are characterized by the presence of a single anchorperson.

Table 1. Composition of the databases used in this paper and in [3]

Paper	Total length (hh:mm:ss)	Number of videos	Number of Broadcasters	Number of Anchor/News-report shots
This	16:54:09	45	2	673 / 8922
[3]	05:05:17	14	2	253 / 3654

When dealing with unbalanced data sets, like in this case, where the anchor shot samples outnumber news report shot samples, the performance is typically reported in terms of *Precision* and *Recall*. However, in order to compare the two algorithms a single figure of merit should be used. In particular, we used the parameter F defined in [9], which combines *Precision* and *Recall* as in the following:

$$F = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

As pointed out in Sect. 2, the original algorithm is characterized by the two thresholds λ and γ . Then, different operating points can be obtained in a *Precision-Recall* plane. Differently, the modified version of the algorithm requires to select only the value of the threshold δ on the lifetime value, since the thresholds λ and γ are automatically calculated. As regards the original version of the algorithm, 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 algorithm, since such maximization should be done on a different set of news videos. It is also worth noting that, as experimentally demonstrated in [10], the choice of the operating point is crucial for the algorithm, as its performance dramatically depends on the choice of the thresholds λ and γ . On the contrary, we experimentally verified that the value of δ can change in a wide range without affecting the performance of the modified algorithm. In particular, it depends only on the length of the news edition, and can be fixed to 2 m for all the news editions shorter than fifteen minutes and to a value equal to 4 m otherwise. Therefore, our algorithm can be really considered as fully unsupervised.

Table 2. The performance of the original algorithm by Gao and Tang and the modified algorithm proposed in this paper in terms of *Precision*, *Recall* and F

	RAI 1			CANALE 5		
	<i>Recall</i>	<i>Precision</i>	F	<i>Recall</i>	<i>Precision</i>	F
Original	0.854	0.947	0.898	0.812	0.681	0.741
Modified	0.923	0.994	0.957	0.852	0.930	0.889

Table 2 reports the performance of each algorithm for the two considered TV-networks. For the news videos of both the TV-networks there is a noteworthy improvement in terms of F . This represents a really valuable result if we consider that the algorithm in [3] is already characterized by a good performance.

Table 3. The performance of the original algorithm by Gao and Tang and the modified version using *i*) only the lifetime control, *ii*) only the automatic threshold computation and *iii*) both the proposed modifications

	RAI 1			CANALE 5		
	<i>Recall</i>	<i>Precision</i>	F	<i>Recall</i>	<i>Precision</i>	F
Original	0.854	0.947	0.898	0.812	0.681	0.741
Original + Lifetime	0.886	0.985	0.933	0.840	0.930	0.884
Original + Automatic Thresholds	0.928	0.923	0.926	0.852	0.761	0.804
Both Modifications	0.923	0.994	0.957	0.852	0.930	0.889

In order to better understand the contribute to the performance improvement given by the introduction of the cluster lifetime and the automatic threshold computation, we calculated the value of the *Precision*, *Recall* and F on the two datasets by

considering the original algorithm by Gao and Tang using only the lifetime control or only the automatic threshold computation. The obtained results are reported in Table 3. From the results reported in Table 3 it is evident that both the proposed modifications allow us to improve the performance with respect to the original algorithm. In particular, the lifetime control allows the proposed algorithm to significantly outperform the original algorithm of Gao and Tang in terms of *Precision*. On the other hand, the automatic selection of the thresholds permits a significant improvement in terms of *Recall*, especially for the RAI 1 videos.

4 Conclusions

In this paper an improved algorithm for anchor shot detection was presented. It was tested on a news video database consisting of about 17 hours, giving rise to a significant improvement with respect to its original version.

In order to further improve the *Recall* value obtained by the proposed algorithm, other information sources, such as the audio track of the news video, could be used. Moreover, other similarity measures between key-frames could be employed for dealing with errors due to the fact that some anchorperson models appear only once in a whole news program. All these matters will be subjects of future investigations.

References

1. B. Furht, S.W. Smoliar, H. Zhang, Video and Image Processing in Multimedia Systems, Kluwer Publishers, Boston (MA), 1996.
2. Y. Avrithis, N. Tsapatsoulis, S. Kollias, "Broadcast news parsing using visual cues: A robust face detection approach", Proc. of IEEE ICME, vol. 3, pp. 1469–1472, 2000.
3. X. Gao, X. Tang, "Unsupervised Video-Shot Segmentation and Model-Free Anchorperson Detection for News Video Story Parsing", IEEE Trans. on Circuits and Systems for Video Technology, vol. 12, no. 9, pp. 765–776, 2002.
4. M. Bertini, A. Del Bimbo, P. Pala, "Content-based indexing and retrieval of TV News", Pattern Recognition Letters, vol. 22, pp. 503–516, 2001.
5. A. Hanjalic, R.L. Legendijk, 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.
6. X. Gao, J. Li, B. Yang, "A Graph-Theoretical Clustering based Anchor Shot Detection for News Video Indexing", Proc. of Int. Conf. on Comput. Intell. and Multimedia App., 2003.
7. 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.
8. R. Balakrishnan, K. Ranganathan, A Textbook of Graph Theory. New York: Springer-Verlag, 1999.
9. L. Chaisorn, T.-S. Chua, C.-H. Lee, "A Multi-Modal Approach to Story Segmentation for News Video", World Wide Web, vol. 6, pp. 187–208, 2003.
10. 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.