A Comparison of Unsupervised Shot Classification Algorithms for News Video Segmentation

Massimo De Santo[1], Gennaro Percannella[1], Carlo Sansone[2], and Mario Vento[1]

[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. Automatic classification of shots extracted by news video plays an important role in the context of news video segmentation. In spite of the efforts of the researchers involved in this field, a definite solution for the shot classification problem does not yet exist. Moreover, the authors of each novel algorithm usually provide results supporting the claim that their method performs well on a set of news videos, without facing the problem of making a wide comparison with other algorithms in terms of key performance indexes.
In this paper, we present an experimental comparison of three shot classification algorithms. We considered only techniques that do not require the explicit definition of a model of the specific news video. In such a way the obtained performance should be quite independent of the news program's style. For testing the selected algorithms, we built up a database significantly wider than those typically used in the field.

1 Introduction

In order to allow a faster and more appealing use of news video databases, indexing and retrieval are essential issues to be addressed. A first step towards an effective indexing is the segmentation of a news video. It implies, at a first stage, the partition of the video into sequences of frames, called *shots*, obtained by detecting transitions that are typically associated to camera changes. Once the shots have been individuated, they can be classified on the basis of their content. Two different classes are typically considered, *anchor shot* and *news report shot* classes. Then, a news video can be segmented into *stories*; each story is obtained by linking a given anchor shot together with all successive news report shots, until another anchor shot occurs.

In this paper, we address the shot classification problem. In the literature, most of the approaches that exploit the video source information use 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.

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 233–241, 2004.

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. A different approach based on the frame statistics is presented in [3], where the authors use a Hidden Markov model (HMM) to classify frames. The features used are the difference image between frames, the average frame color and also the audio signal. In this case the HMM parameters are evaluated during a training phase.

Finally, some authors [4,5,6] propose methods that are substantially unsupervised and do not require the explicit definition of an anchor shot model. In particular, in [4] a graph-theoretical cluster analysis method is employed. As pointed out by the authors, this approach fails when identical or very similar news-report shots appear in different stories of the same news program. In [5] shot classification is firstly performed on the basis of a statistical approach and then refined by considering motion features. Here the authors assume that in an anchor shot both the camera and the anchor are almost motionless. In our opinion, however, this hypothesis is not completely acceptable. In [6] a template-based method is proposed. The template is found in an unsupervised way and it does not depend on a particular threshold. However, the authors assume that different anchor shot models share the same background. This is not true for most news stations: because of different camera angles, different models can have different backgrounds.

From the previous analysis it appears evident that a definite solution for the shot classification problem does not yet exist, since it is always possible to find a case in which the assumptions of a given technique fail. Moreover, the authors of each algorithm usually provide results supporting the claim that their method performs well on a set of news videos, without facing the problem of making an extensive comparison with other algorithms in terms of key performance indices.

Starting from these considerations, in this paper we present an experimental comparison of three shot classification algorithms among those presented so far in the literature. We have chosen to consider only techniques that do not require the explicit definition of a specific model of the anchorperson shot. In such a way the obtained performance should be quite independent of the specific style of the news program.

In order to test the selected algorithms in a significant way, we built-up a database that is twice the biggest database reported until now in the scientific literature [4]. Namely, we used a news video database consisting of about 10 hours with 464 anchor shots and 5705 news report shots.

The organization of the paper is as follows: in section 2 the three selected shot classification algorithms are presented. In section 3 the database used is reported together with the tests carried out in order to assess the performance of the selected algorithms. Finally, in section 4, some conclusions are drawn.

2 The Selected Algorithms

Among all the techniques described in the introduction, we choose to consider for comparison only those that neither use a model-based approach nor require a specific training phase. This choice is justified by the consideration that the variety of existing news programs in the world is so high to make it practically infeasible building a good general model. Even when limiting the attention to a single broadcaster, as it could be reasonable in a real application, the news model can repeatedly change over

the time. In this case, a model-based approach could provide good results only for a small period of time, while requiring a continuous re-modeling to guarantee acceptable performance for a longer period. For the same reason, methods requiring a specific training phase are not very suitable for the application at hand, given that it would be necessary re-training the system for each change in the news style.

In particular, on the basis of the best results available in the literature, we considered the shot classification algorithms proposed by Bertini *et al*. in [5], Gao and Tang in [4] and Hanjalic *et al.* in [6]. Hereinafter, for the sake of simplicity, we will refer to these algorithms with the terms BER, GAO and HAN, according to the first three letters of the first author.

In the following we will briefly recall the rationale inspiring these algorithms.

2.1 BER Algorithm

The shot classification is here performed on the basis of a statistical approach and of the motion features of the anchor shots, without requiring any model [5]. The statistical approach is based on the consideration that anchor shots are repeated at variable length throughout the video. The obtained classification is successively refined by considering also motion features: according to the authors, it is reasonable to assume that in an anchor shot both the camera and the anchorperson are almost motionless. More in details, the first step of the process is based on the computation, for each video shot S_k, of the so-called *shot lifetime* $L(S_k)$. It measures the shortest temporal interval that includes all the occurrences of shots with similar visual content within the video and is used to perform a first shot classification. In this case, for each video shot the first frame is chosen as key-frame. Having defined a suitable similarity measure between two frames, the similarity between two shots is evaluated as the similarity between their key-frames. By indicating with t_i the value of a time variable corresponding to the occurrence of the key-frame of the shot S_i, we can build, for each shot S_k, the set T_k. It contains all the values t_i relative to the shots S_i whose similarity with the shot S_k is lower than a suitable threshold τ_S; in other words, it contains the time occurrences of all the shots similar to S_k. The shot lifetime $L(S_k)$ can be then defined as the difference between the last and the first value of t_i

Since anchor shots occur repeatedly through the video, the shot classification is performed by attributing to the anchor shot class all the shots S_k having the value of $L(S_k)$ greater than a suitably chosen threshold τ_l. The value of τ_l can be determined according to the statistics of the specific video database. In [5] it is fixed to 4.5 s.

This classification is then refined by computing an index Q_S that measures the quantity of motion for each candidate anchorperson shot. This index Q_S is calculated as the sum of all the frame-to-frame difference between the key-frame and all the subsequent frames in the shot. So, only those shots whose Q_S value does not exceed a threshold τ_Q are definitely classified as anchor shots.

Note that for this method three thresholds need to be evaluated: τ_S, τ_l, and τ_Q.

2.2 GAO Algorithm

In this case [4], video shots are classified by using an algorithm based on graph-theoretical cluster (GTC) analysis. More in details, the authors 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 s, since this shot should involve at least one sentence pronounced by the reporter. Therefore, if a shot lasts less than 2 s 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 constructs the minimum spanning tree (MST) on these vertices. To do that, a distance between key-frames, based on the color histograms, is defined. It is used for weighting each edge connecting two vertices. Successively, by removing from the MST all the edges 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 (i.e., 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 color histograms but different content. To detect this situation, a spatial difference metric (SDM) between two key-frames is proposed. If a cluster has an average SDM higher than a threshold λ, the whole cluster is removed from the anchorperson frame list.

It is worth noticing that, in this case, two thresholds, γ for the GTC algorithm and λ for the post-processing step, need to be specified in advance.

2.3 HAN Algorithm

In [6] a template-based method is proposed. It is based on the assumption that an anchor person shot is the only shot that has multiple match of most of its visual content along the whole video, and consists of two steps: a unsupervised procedure for finding the template shots and its use to detect all the anchor person shots in the video sequence by applying an adaptive thresholding.

The authors assume that the first anchor shot in a news video appears within the first N shots (in the paper N is fixed to 5). A dissimilarity measure is defined between two shot, as it will be specified later. Each shot S_k with $k \in [1,N]$ is then matched with all the other shots, obtaining a set of dissimilarity values for each S_k. For each S_k the P best matches (i.e., the lowest values) out of its set of dissimilarities are averaged to compute the overall matching value of the shot S_k. The shot with the lowest overall matching value is assumed to be an anchor shot and is used as template.

The dissimilarity measure between two shots is defined as follows: each shot is represented by means of two frames, one close to the beginning of the shot and the other close to the end of the shot. These frames are merged into the so-called *shot image*. Each shot image is divided into blocks of $M1 \times M2$ pixels and a distance in the

*L*u*v* color space between two blocks belonging to different shot images is defined. The dissimilarity measure between two shot images S_i and S_j, is then defined as the minimum value among all the average distances obtainable by considering each possible matching of blocks b_k belonging to S_i and blocks b_h belonging to S_j. In the paper, for minimizing the computational effort of the exact evaluation of this measure, only the *C* blocks more similar each other are considered.

Once the anchor shot template has been found, all the remaining shots are checked for individuating the other anchor shots. In particular, all the shots whose similarity with the template shot is lower than an adaptive threshold *M* are detected as anchor shots. Such threshold is proportional to a suitably chosen parameter *w*.

It is worth noticing that the values of the parameters *M*1, *M*2, *P*, *C* and *w* need to be fixed for this algorithm. In [6], for two video sequences with key-frames of size 165x144 and 180x144 respectively, *M*1 and *M*2 were both fixed to 8, while *P* was set to 3, *w* to 3.0 and *C* was considered as the 70% of the total number of blocks.

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.

So, we decided to build-up a new database. The acquisition was performed by means of the digital satellite decoder *emme esse 6000pvr*. It has an internal hard disk that allowed us to record in the DVB MPEG-2 format. Then, the videos were transferred on a PC, so preserving the broadcasting quality. We encoded the videos in the MPEG-1 format using the TMPGEnc encoder (ver. 2.01). The parameters used to encode the videos were selected taking into account the storage requisites, without decreasing the performance of the algorithms with respect to those obtainable with the original full-quality videos. In particular, we selected four videos from our database for analyzing the dependence of the algorithms' performance on both the frame size and the bit-rate. Two frame sizes have been considered: 704x576 and 352x288. We verified that the information loss due to the 352x288 frame size does not allow the algorithms to perform the same as on the full-quality videos. On the other hand, if the video is coded with at least 1500 kbit/s and the frame size is 704x576, all the algorithms perform as well as on the original videos. In order to keep a tolerance margin, we decided to encode the videos with approximately 2000 kbit/s.

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. However, while in the first case the different editions are usually less than ten, in the latter the number of different models is significantly large. As a consequence, we preferred to test our system on a database composed by all the different news videos captured from a single broadcaster, rather than perform tests with only few samples belonging to a large number of different broadcasters. Even if some archiving companies work with large quantities of videos

from different sources, this approach fits the most realistic use case for the proposed system. A typical broadcaster, in fact, should be interested to employ such a system for analyzing all the editions of its news videos.

The database used in this paper is composed by more than thirty news videos from the main Italian public TV-network (namely, RAI 1). As it can be easily noted from Table 1 its size is large; this is more evident if it is compared with the databases used in the papers of BER, GAO and HAN.

Table 1. Composition of the databases used in this paper and in [4], [5] and [6].

Paper	Total length (hh:mm:ss)	Number of videos	Number of Broadcasters	Number of Anchor/News-report shots
This	09:24:19	34	1	464 / 5705
[4]	05:05:17	14	2	253 / 3654
[5]	02:41:00	12	6	66 / 665
[6]	00:37:00	2	-	22 / --

As a first step for the assessment of the performance of the three anchor shot classification algorithms, we calculated their *Precision-Recall* curves [7]. Each point of these curves represents the performance in terms of *Precision* and *Recall* obtained by the algorithm using a specified set of thresholds. Each considered technique is characterized by several thresholds; hence, for each technique a family of *Precision-Recall* curves can be drawn. Each curve is obtained by varying the value of a threshold, holding fixed the remaining ones. In particular, in Figures 1-3 the operating curves for GAO, BER and HAN, respectively, are shown.

The operating curves for GAO were obtained by varying both the thresholds λ (in the range 1500-3500 with step 500) and γ(in the range 20000-45000 with step 5000). Differently, the curves for BER were obtained by varying the value of τ_S in the range 25000-50000 with step 5000 and τ_Q in the range 25-150 with step 25, while holding fixed $\tau_l = 4.5$. This threshold, in fact, does not significantly influence the overall algorithm performance. Finally, there is a single operating curve for HAN. It is obtained by varying the value of w (in the range 1.5-4.0 with step 0.1), while holding fixed the other parameters. We used the same values suggested by the authors for N, P and C, while $M1$ and $M2$ were both fixed to 32, so as to have the same number of blocks per shot image of the original paper. In this case it is founded that variations in the values of P and C do not implies variations on the operating curve for HAO.

So, a first result is that for BER and HAN only a subset of the parameters are really significant. In fact, it is possible to allow some parameters to vary within wide ranges of values without significant changes in the obtainable performance. However, the families of curves shown in Figs. 1-3 are not suitable for a comparison of the three algorithms. Depending on the chosen operating point, it is possible to detect a different curve which allows maximizing the performance. This also implies that a unique set of values that maximizes the performance of each algorithm does not exist, except that for the HAN algorithm. Hence, the comparison of the three algorithms has been carried out on the basis of the envelope of the *Precision-Recall* curves. For each algorithm the points of the envelope were obtained by considering for each value of the *Precision* the values of the parameters that maximized the *Recall*.

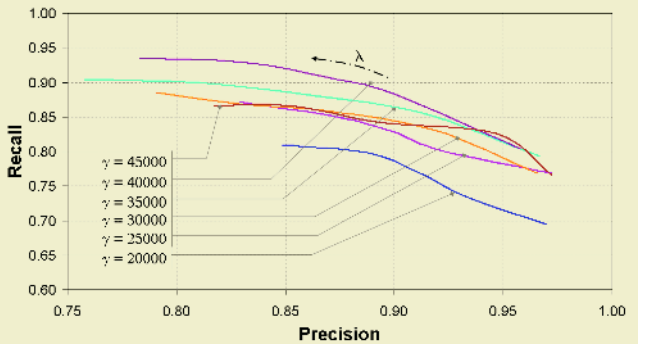

Fig. 1. Operating curves for GAO. Each curve is obtained by varying the value of λ while holding fixed the value of γ. The dashed arrow indicates the direction of increasing values of λ. The six curves accounts for six different values of the threshold γ.

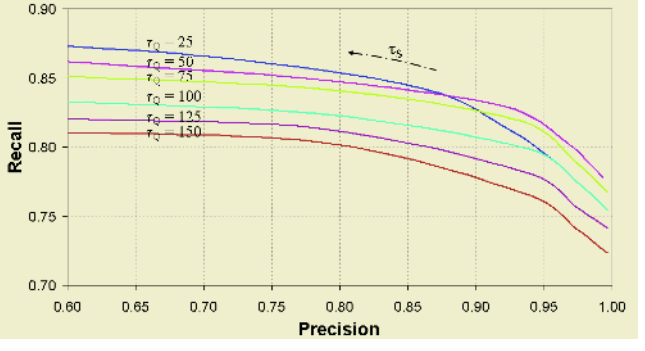

Fig. 2. Operating curves for BER. Each curve is obtained by varying the value of τ_S while fixing the other parameters' values. The dashed arrow indicates the direction of increasing values of τ_S. The eight curves are relative to eight different values of the threshold τ_Q.

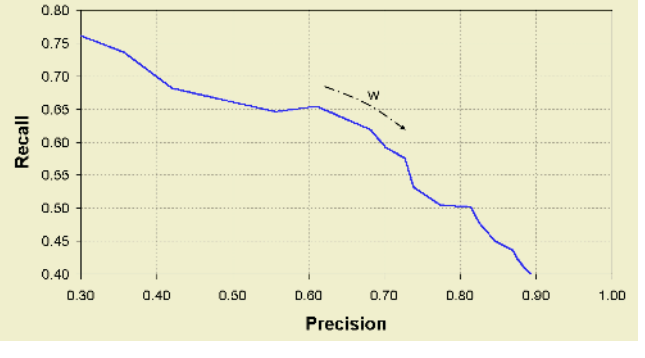

Fig. 3. Operating curve for HAN. It is obtained by varying the value of *w* while holding fixed the other parameters' values. The dashed arrow indicates the direction of increasing values of *w*.

In Figure 4 the envelopes of the operating curves of the three algorithms are reported. Curves in Figure 4 clearly show that HAN performs much worse than BER and GAO for all the operating conditions. Differently, the behaviour exhibited by both the other two algorithms is characterized by a high and stable value of the *Recall* (above 0.80) for almost all the values of the *Precision*. Furthermore, it is interesting to note that the GAO algorithm is preferable if we are looking for high *Recall* values (i.e., higher than 0.90). On the contrary, if it is mandatory to have very high *Precision* values (i.e., almost no false alarms), the BER algorithm must be chosen.

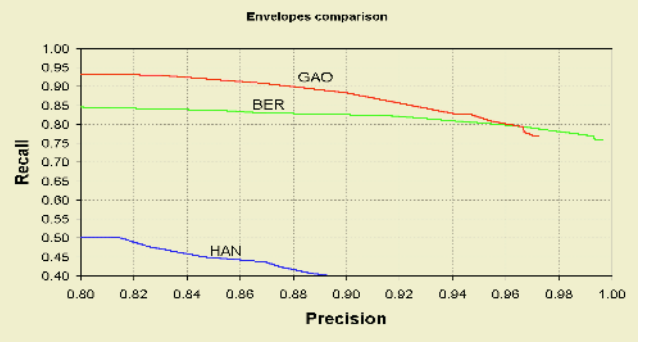

Fig. 4. Envelopes of the operating curves for the three considered algorithms.

In order to provide a more global comparison among the three different algorithms and for comparing the results obtained on our database with those achieved by the algorithms in their original papers, a unique figure of merit, as the parameter F defined in [8], can be used. It combines *Precision* and *Recall* as in the following:

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

For choosing the operating conditions of BER and GAO algorithms to be used on the whole database, a preliminary tuning phase was required. In particular, we chose the optimal values of the thresholds by means of an empirical optimization, i.e., by maximizing F over a predefined set of videos. In this set we included the same videos already used for setting up the MPEG-1 coding parameters for the whole database. These videos were not included in the successive tests.

Table 2 reports the global performance of each algorithm; a first result is the discrepancy between the performance reported in Table 2 and the results presented in the original papers (reported for the sake of comparison in parenthesis). All the algorithms perform worse on our database: this is particularly true for the HAN algorithm. Moreover, GAO algorithm performs better in terms of *Precision*, while in [4] its behavior was the opposite one. Analogously, BER algorithm exhibits on our database a *Recall* value higher than the *Precision* one, differently from the behavior described in [5]. Since the anchor shots/news-report shots ratio of our database is quite similar to those of the databases used in the original papers (excluding HAN, where the ratio is not specified), such discordances are mainly due to the different size of the database used for the testing the algorithms in this paper. This confirms the importance of testing algorithms on a large and significant database. Finally, the results reported in Table 2 point out that, even if BER algorithm outperforms the remaining two in terms of *Precision* and F, GAO exhibits the best value of the *Recall*.

Table 2. The performance of the three considered algorithm in terms of *Precision*, *Recall* and *F*. For the sake of comparison, the results obtained in the original papers are reported in parenthesis.

	Recall	***Precision***	***F***
GAO	0.929 (*0.973*)	0.842 (*0.976*)	0.881 (*0.974*)
BER	0.816 (*0.970*)	0.987 (*0.955*)	0.892 (*0.962*)
HAN	0.623 (*1.000*)	0.692 (*0.917*)	0.655 (*0.957*)

4 Conclusions

In this paper an experimental comparison of three unsupervised algorithms for anchor shot classification was presented. The comparison has been carried out on a news video database consisting of about 10 hours with 464 anchor shots and 5705 news report shots. As it could be expected, it does not exist an algorithm that is definitively better than the others. While HAN algorithm performs always the worst, BER algorithm typically outperforms the remaining two in terms of *Precision*, while GAO exhibits in general higher values of *Recall*. However, the choice of the most suitable algorithm at hand strongly depends on the selected operating conditions.

Future steps of this activity will involve the comparison of other anchor shot detection algorithms. We are also planning to investigate in more details how the similarity measure between key-frames influences the algorithms' performance. To this aim, other similarity measures will be also considered.

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. IEEE Intern. Conf. on Multimedia and Expo, vol. 3, pp. 1469–1472, 2000.
3. S. Eickeler, S. Muller, "Content-based video indexing of TV broadcast news using Hidden Markov Models", Proc. IEEE International Conference on ASSP, pp. 2997-3000, 1999.
4. 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.
5. M. Bertini, A. Del Bimbo, P. Pala, "Content-based indexing and retrieval of TV News", Pattern Recognition Letters, vol. 22, pp. 503-516, 2001.
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. 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. 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.