Automatic Cut Detection in MPEG Movies: A Multi-expert Approach

Massimo De Santo[1], Gennaro Percannella[1], Carlo Sansone[2], Roberto Santoro[1], 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,rsantoro,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. In this paper we propose a method to detect abrupt shot changes in MPEG coded videos that operates directly on the compressed domain by using a Multi-Expert approach. Generally, costly analysis for addressing the weakness of a single expert for abrupt shot change detection and the consequent modifications would produce only slight performance improvements. Hence, after a careful analysis of the scientific literature, we selected three techniques for cut detection, which extract complementary features and operate directly in the compressed domain. Then, we combined them into different kinds of Multi-Expert Systems (MES) employing three combination rules: Majority Voting, Weighted Voting and Bayesian rule. In order to assess the performance of the proposed MES, we built up a huge database, much wider than those used in the field. Experimental results demonstrate that the proposed system performs better than each of the three single algorithms.

1 Introduction

Information filtering, browsing, searching and retrieval are essential issues to be addressed in order to allow a faster and more appealing use of video databases. Even if it does not yet exist a unique and definitive solution to the former problem, the field experts have agreed upon a position: the first step toward an effective organization of the information in video databases consists in the segmentation of the video footage in shots, defined as the set of frames obtained through a continuous camera recording. There are two different types of transitions between shots: abrupt and gradual. The difference between them relies on the number of frames involved, which are two in the case of abrupt shot changes and more than two in the case of gradual shot changes. In the latter case, different types of shot transitions may be outlined as dissolves, fades, wipes, iris, etc., according to the mathematical model used to transform the visual content from a shot to the successive one. The automatic detection of gradual transitions is much more complicated than that of abrupt shot

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 699-708, 2002.

changes and requires more complex mathematical models. However, gradual transitions are also less frequent than abrupt shot changes; therefore, in this paper we focused our investigation only on abrupt shot changes detection. It is worth to consider that automatic abrupt shot changes detection (SCD) is not a trivial task and is often complicated by some video effects, like camera or objects movements, impulsive variations of luminance signals, that may be easily confused with abrupt shot changes.

It has to be noted that video sources are often provided in compressed form according to standards like MPEG. In the recent past, many researchers have tried to face the problem of cut detection by processing videos in compressed form. In fact, the direct analysis in the coded domain offers at least two advantages: firstly, the computational efficiency of the whole process is improved; secondly, video compression is generally performed using signal processing techniques capable of deriving features for video segmentation, e.g. motion vectors in MPEG coding. Thus, such features become readily available for any parsing operation, and would have to be re-derived if a decoding step were applied. For these reasons, we perform the whole analysis of the video stream directly in the MPEG coded domain.

In the scientific literature many techniques for SCD have been proposed. However, the efforts for increasing performance of a single classifier appear, in general, unjustified, especially when the classifier has been repeatedly improved over the time, by adjusting its features, the learning procedures, the classification strategies and so on. Generally, costly analysis for addressing the weakness of a single classifier and the consequent modifications would produce only slight performance improvements. In these cases, the ensemble of rather simple experts, complementary as regards their errors, makes it possible to improve the overall performance, and often relatively little efforts are rewarded by high performance increases.

Therefore, our intention was to employ a Multi-Expert approach that can give good performance improvements with relatively few efforts. The use of a Multi-Expert System (MES) for complex classification tasks has been widely explored in the last ten years [1, 2]. The underlying idea of using a MES is to combine a set of experts in a system taking the final classification decision on the basis of the classification results provided by any of the experts involved. The rationale of this approach lies on the assumption that the performance obtained by suitably combining the results of a set of experts is better than that of any single expert. The successful implementation of a MES requires experts which use complementary features, and the definition of a combining rule for determining the most likely class a sample should be assigned to, given the class it is attributed to by each single expert.

Hence, our idea was to select three methods proposed in the scientific literature and to combine them into a MES according to the most commonly used combining rules: Majority Voting, Weighted Voting and Bayesian rules [1, 2]. We considered two principal aspects when choosing the algorithms to integrate in our MES: the complementarity of the used features and the performance declared by the authors.

For the training and testing phases, we used a database consisting of more than 130 thousands frames with a percentage of about 1% of abrupt cut frames. This is a significant amount of both frames and cuts, especially if compared to the size of the data sets usually employed in this scientific realm. The experimental results showed that the proposed MES performs better than each of the considered classifier.

The organization of the paper is the following: in section 2 we report about previous works in the field of automatic abrupt cut detection in the MPEG compressed domain; in section 3 the proposed system architecture is presented, together with some details about the cut detection algorithm implemented into the three experts; in section 4 we analyze the experimental campaign carried out in order to assess the performance of the proposed system; finally, in section 5 we draw conclusions and discuss on the future work.

2 Previous Work

In this section, we briefly report about proposed methods for automatic abrupt cut detection, which, according to our opinion, are the most representative. As mentioned in the introduction, we focus our attention on MPEG coded videos. A possible taxonomy for classifying algorithms for automatic detection of video shot transitions can be based on the required level of decoding. From this point of view, we devised four different groups of techniques sorted according to the increasing number of decoding steps required to derive the basic information needed for shot boundaries detection: bit rate, macroblock prediction type, motion vectors, DCT coefficients.

Bit rate techniques [3, 4] rely on the idea that large variation in the visual content between two consecutive frames generally results in a large variation in the amount of bits used for coding the DCT coefficients of the blocks of the respective frames. Anyway, the variations in the amount of bits may occur both when a cut or other effects like zooming, panning or dissolves are present. Being the used information trivial, there are no ways to distinguish among these cases. The idea behind macroblock prediction type techniques for SCD [5, 6] is that the visual change generated by a cut usually gives rise to specific patterns of macroblocks into the frames across the shot boundary. Therefore, recognizing these patterns means recognizing cuts. The use of features based on motion vectors [7] relies on a very simple principle: temporally adjacent frames belonging to the same shot are usually characterized by the same motion. Hence, the motion vectors in the inter-coded frames (i.e. *P* or *B*) might be used to this aim. In particular, the difference between the motion vectors of a block of two successive inter-coded frames should be small or large if the two frames are respectively in the same shot or not.Another source of information that has been often used for shot segmentation is represented by DCT coefficients [8, 9, 10, 11]. The idea is that a variation in the content of a block of 8x8 pixels results in a variation in the content of the block in the transformed domain, and so in its DCT coefficients.

Generally speaking, all the above mentioned techniques for SCD operate in a similar fashion. Each one parses a MPEG coded video frame by frame, computes the distance between each couple of successive frames and, if this difference is greater or lower than a fixed threshold, they declare or not a cut between the two frames the difference is referred to. The distinction among these techniques relies on the way they compute the difference between two frames. Therefore, each technique for SCD can be viewed as a single classifier that for each couple of frames declares the presence of a cut or not.

3 The Proposed System Architecture

According to the rationale inspiring MES, we selected three techniques for SCD (three experts) whose features were complementary and combined them according to a parallel scheme, as shown in Fig. 1. Each of the three single classifiers receives in input the MPEG coded bitstream and for each couple of frames provides its own classification (e.g. cut or not cut). Then the combination module of the MES, for each couple of frames, provides the final classification on the basis of the outputs of the three experts and of the implemented combining rule.

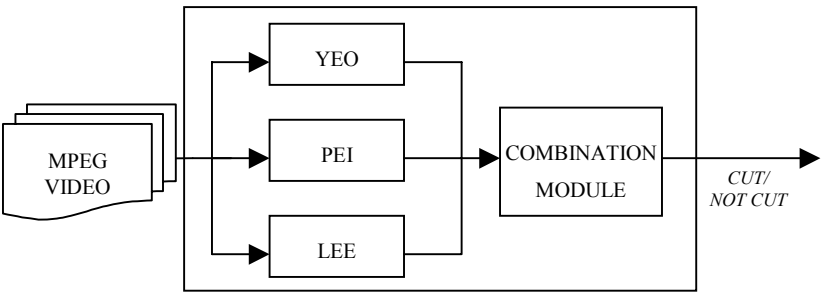

Fig. 1. The system architecture of the proposed MES

The experts used in our system implement the SCD algorithms proposed by Yeo et al. in [8], Pei et al. in [5] and Lee et al. in [9]. These techniques offers the advantage to extract complementary features (e.g. DC-image, edges and macroblock prediction type) and the performances reported by the authors are interesting, as shown in Table 1. Hereinafter, for the sake of simplicity we will refer to these three classifiers with the terms YEO, PEI and LEE, according to the name of the first author. The performance index (*PI*) used for comparing the various techniques is defined by the sum of *Precision* and *Recall*. In (1), (2) and (3), we reported the formulas of *Precision*, *Recall* and *PI*, respectively:

$$Precision = \frac{cd}{cd + f} \tag{1}$$

$$Recall = \frac{cd}{cd + m} \tag{2}$$

$$PI = Precision + Recall \tag{3}$$

where *cd* is the number of correctly detected cut, *f* is the number of false positive and *m* is the number of misses.

The SCD method proposed by Yeo et al. employs the average value of the luminance computed on each frame. For each video frame a *DC-image* is constructed; such image is obtained considering for each 8x8 pixels luminance block only the value of the DCT-DC coefficient. Therefore, a frame of 352x288 pixels is represented through 44x36 DC coefficients. For each couple of successive frames, YEO computes the distance as the sum of the absolute differences among the corresponding pixels of

the two *DC-images*. Then, it considers sliding windows of *m* frames, computes *X* and *Y*, respectively as the first and the second maximum distances between each couple of successive frames into the window: if *X* is *n* times greater than *Y*, then a cut is declared between the two frames whose distance is *X*.

Table 1. Experimental results reported by Yeo et al., Pei et al. and Lee et al.

	Precision	Recall	PI
YEO	0.93	0.93	1.86
PEI	0.99	1	1.99
LEE	0.97	0.99	1.96

DCT coefficients have been used also in the SCD method proposed by Lee et al. In this case, for each 8x8 pixels luminance block of every frame, seven DCT coefficients are needed. Anyway in [9], the key idea is to perform cut detection on the basis of the variations of the edges. In fact, the authors developed a mathematical model to approximately characterize an eventual edge in the block by using only on the first seven DCT coefficients in the zig-zag order. The characteristics of an edge are represented in terms of its intensity (*strength*) and orientation (ϑ). The technique works as follows: for each frame, they compute the histogram of the edge strengths, *H(strength)*, and the histogram of the edge orientations, $H(\vartheta)$. Then, for each couple of successive frames, they compute the differences *D(strength)* and $D(\vartheta)$ between their *H(s)* and $H(\vartheta)$, respectively. Finally, the interframe distance is obtained as a linear combination of *D(s)* and $D(\vartheta)$; if such distance is higher than a fixed threshold, a cut is declared between the two frames.

The third technique for SCD we used is that proposed by Pei et al. In this case, the employed feature is the macroblock prediction type. Each macroblock consists of four 8x8 pixels blocks and can be coded as *I*, *P* or *B*. The idea is that a particular pattern of coded macroblocks should reveal the presence of a cut. As an example, typically most macroblocks of a *B* frame are coded referring both to a precedent anchor frame (*I* or *P*) and to a successive anchor frame (i.e. *forward and backward prediction*). Differently, when a cut occurs between a *B* and a successive anchor frame, most macroblocks of a *B* frame are coded only referring to the previous anchor frame (i.e. *forward prediction*). This SCD technique is very fast as the macroblock coding type is a readily available information, requiring only few MPEG decoding steps.

As regards the combiner, we implemented the most common combination rules: Majority Voting, Weighted Voting and Bayesian rule. In particular, for the Weighted Voting and the Bayesian rules, we considered the combinations of all the three algorithms and the three possible combinations of two out of three algorithms. Therefore, on the whole, we considered nine combinations.

As regards the weighted voting, it is worth to specify that the votes provided by each expert have been weighted proportionally to the percentage of correct recognition evaluated on the training set for each class. As an example, if a test set sample has been classified as a cut by the *i*-th expert, this vote will weight 0.8 if the percentage of correctly detected cut on the training set for the *i*-th classifier was 80%. As to the Bayesian rule, the set up was done according to [12].

4 Experimental Results

In this section we report the description of the database used for the experimental campaign, together with experimental results provided by the single experts and the proposed MES.

4.1 The Video Database

In the field of SCD a topic still in progress is the definition of a common database. Nowadays, each researcher involved in this field uses a different database, so making comparisons among SCD algorithms a very complicated task. Moreover, quite often, these databases consist of too few frames and cuts.

In order to carry out a significant analysis of the proposed system, we set up a database consisting of sixteen excerpts of MPEG coded movies of various genres for a total amount of 134314 frames and 1071 cuts. This is a significant amount of frames above all if compared to other databases used in the SCD field. As example, in Table 2, we reported a comparison among our database and databases used in [5, 8, 9].

Table 2. Size comparison of the databases used in [5, 8, 9] and in this paper

Databases	Number of frames	Number of cuts
Yeo et al. [8]	9900	41
Pei et al. [5]	36000	269
Lee et al. [9]	80887	611
This paper	134314	1071

With the aim to obtain reliable experimental results, we selected excerpts of videos MPEG of various genres. Furthermore, the database contains several common video effects as zooming, panning, dissolves and fades: this is important as it allows stressing the algorithms in several difficult conditions. Moreover, in the database is also present the ground truth, therefore it is available, for each frame, the information concerning the presence, or the absence, of a cut. This information is structured as follows: if the frame *i* is labeled as a cut, it is intended that there is a cut between the frame *i* and the frame *i*-1. This information is essential to assess the performance of every employed classifier.

The whole database has been divided into two sets (training and test, respectively) of approximately the same size: both sets contain eight video fragments. In Table 3 there are reported some details about the composition of the two sets.

Table 3. Composition of the training and test sets

	Number of frames	Number of cuts
Training Set	64343	543
Test Set	69971	528

The training set was used to fix the optimal threshold for each classifier. To this aim, the optimal threshold is intended as the threshold that allows the classifiers to maximize *PI* (the index defined in Section 2) on the training set.

4.2 Experimental Results

In this section we report the performance of the three single classifiers and of the proposed MES, evaluated on the test set.

Firstly, we compared the performance obtained by the three single classifiers. From Table 4, it is possible to note that performances obtained by YEO and PEI are much better than those obtained by LEE. This is due to the very high percentage of false cuts detected by LEE, which degrades the value of *Precision* and consequently the performance index. As shown in Table 4, the performance index for PEI is 1.83 versus 1.70 for YEO and 0.66 for LEE.

Table 4. A comparison among performances of the three single classifiers on the test set

Expert	%cd	%f	Precision	Recall	PI
YEO	82.00	0.08	0.88	0.82	1.70
PEI	91.60	0.07	0.91	0.92	1.83
LEE	64.40	23.50	0.02	0.64	0.66

Table 5 reports the coverage Table evaluated on the test set for the combination of the YEO, PEI and LEE experts, in terms of correctly detected cut and false positive. Table 5 is extremely useful as it allows sketching the complementarity of the employed experts. As an example, the value of *%cd* for "None" represents the percentage of correctly detected cuts which neither YEO nor PEI nor LEE are able to detect. Therefore there is no MES that by combining the three single classifiers can detect these cuts, independently from the employed combination rule. Obviously, the complement of this percentage constitutes the upper bound for the percentage of correctly detected cuts that any MES can obtain by combining YEO, PEI and LEE.

Table 5. Coverage table evaluated on the test set for the combination of the YEO, PEI and LEE experts, in terms of correctly detected cut and false positive

Number of classifiers	Classifiers	%cd		%f	
		Percentage	Sum	Percentage	Sum
3	YEO, PEI and LEE	52.65	52.65	0	0
2	YEO and PEI	27.84	35.80	0	0.06
	YEO and LEE	1.33		0.03	
	PEI and LEE	6.63		0.03	
1	Only YEO	0.19	8.52	0.05	22.11
	Only PEI	4.54		0.04	
	Only LEE	3.79		22.02	
0	None	3.03	3.03	77.83	77.83

The theoretic MES that could provide this percentage is called *oracle.* In our MES the *oracle* is characterized by a *%cd* of 96.97. Moreover, from Table 5, it is possible to deduce the performance that majority voting MES is able to gain. The sum of percentages of correctly detected cuts provided by at least two out of three single classifiers is 88.45, which is the percentage of correctly detected cuts by using the majority voting rule. Table 5 shows also the coverage percentages of false positive

obtained by YEO, PEI and LEE on the test set. Here, the value obtained in case of "None" means that on 77.83% of the test set frames, neither YEO nor PEI nor LEE give a false positive. Interestingly enough, the lower bound for the percentage of false positive is 0%; while the percentage of false positive given by using the majority voting rule is 0.06.

In Table 6, a comparison among all considered classification systems is reported, sorted according to the global performance index *PI*. For simplicity we indicated the various MES as follows: MV, W and BAY stand for majority, weighted voting and Bayesian rule respectively; Y, P and L are abbreviations for YEO, PEI and LEE. As an example, BAY-YP is the MES obtained by applying the Bayesian rule to YEO and PEI classifiers.

From table 6, it can be observed that some MES performs exactly the same. This is due partly to the small number of experts (three) and mostly to the small number of classes (two). Table 6 reports in bold the MES which obtained the highest percentage of correctly detected cuts, lower percentage of false positive and higher value of performance index *PI*. In the first row of the same table, we reported also the performance of the *oracle*. Therefore, we can conclude that there are three MES (i.e. BAY, BAY-YP and W-YP) which performs better than the single expert (i.e. YEO, PEI and LEE), considered individually. It is also very interesting to note that the maximum improvement to *PI* that the multi-expert approach would give is 0.17 and the best MES are able to recover about 21% of this maximum improvement; such point constitutes a very good result.

Table 6. Parameters *%cd*, *%f*, *Precision*, *Recall* and *PI*, evaluated on test set, for all the considered classification systems

Algorithms	%cd	%f	Precision	Recall	PI
Oracle	96.97	0	1	0.97	1.97
BAY-YPL	93.20	0.05	0.93	0.93	1.86
BAY-YP	93.20	0.05	0.93	0.93	1.86
W-YP	93.20	0.05	0.93	0.93	1.86
BAY-PL	91.60	0.07	0.91	0.92	1.83
W-PL	91.60	0.07	0.91	0.92	1.83
PEI	91.60	0.07	0.91	0.92	1.83
MV	88.45	0.06	0.92	0.88	1.80
W-YPL	88.45	0.06	0.92	0.88	1.80
BAY-YL	82.00	0.08	0.88	0.82	1.70
W-YL	82.00	0.08	0.88	0.82	1.70
YEO	82.00	0.08	0.88	0.82	1.70
LEE	64.40	23.50	0.02	0.64	0.66

5 Conclusions

Automatic abrupt cut detection in the MPEG coded domain is still an open problem. In spite of efforts of many researchers involved in this field, there are not yet techniques which provide fully satisfactory performances.

In this context, our idea was to employ a Multi-Expert approach to combine some of the best techniques available in the scientific literature with the aim to improve the recognition rates. Therefore, we implemented three algorithms for cut detection, operating directly on the compressed format and combined them in a parallel MES using common combination rules. The experimental results have demonstrated that the bayesian combination of the three algorithms (also only the best two) performs better than each classifier considered individually, with respect to each of the considered performance parameters: percentage of correctly detected cuts, percentage of false positive and value of *Precision* + *Recall*.

At this moment, as future direction of our research we foresee from a side the integration in the MES of other cut detection techniques in order to further improve the overall performances of the system and on the other side the application of the MES approach to the detection of gradual transitions.

References

1. T.K. Ho, J.J. Hull, S.N. Srihari, Decision Combination in Multiple Classifier Systems, IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(1), (1994), 66-75.
2. J. Kittler, M. Hatef, R.P. W. Duin, J. Matas, On Combining Classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), (1998), 226-239.
3. J. Feng, K.T. Lo and H. Mehrpour, "Scene change detection algorithm for MPEG video sequence", Proc. of the IEEE International Conference on Image Processing, vol. 2, pp. 821–824, Sept. 1996.
4. G. Boccignone, M. De Santo, and G. Percannella, "An algorithm for video cut detection in MPEG sequences," Proc. of the IS&T/SPIE International Conference on Storage and Retrieval of Media Databases 2000, pp. 523-530, Jan. 2000, San Jose, CA.
5. S.C. Pei, Y.Z. Chou, Efficient MPEG compressed video analysis using macroblock type information, in IEEE Transactions on Multimedia, 1(4), (1999), 321–333.
6. J. Nang, S. Hong, Y. Ihm, "An efficient video segmentation scheme for MPEG video stream using Macroblock information", Proc. of the ACM International Conference on Multimedia, pp. 23-26, 1999.
7. S.M. Bhandarkar, A.A. Khombhadia, "Motion-based parsing of compressed video", Proc. of the IEEE International Workshop on Multimedia Database Management Systems, pp. 80-87, Aug. 1998.
8. B.L. Yeo, B. Liu, Rapid Scene Analysis on Compressed Video, IEEE Transactions on Circuits and Systems for Video Technology, 5(6), (1995), 533-544.

9. S.W. Lee, Y.M. Kim, S.W. Choi, Fast Scene Change Detection using Direct Features Extraction from MPEG Compressed Videos, IEEE Transactions on Multimedia, 2(4), (2000), 240-254.
10. N.V. Patel, I.K. Sethi, Compressed video processing for cut detection, IEE Proceedings on Vision, Image and Signal Processing, 143(5), (1996), 315–323.
11. S.S. Yu, J.R. Liou, W.C. Chen, Computational similarity based on chromatic barycenter algorithm, IEEE Transactions on Consumer Electronics, 42(2), (1996), 216-220.
12. L. Xu, A. Krzyzak, C.Y. Suen, Methods of Combining Multiple Classifiers and Their Application to Handwritten Numeral Recognition, IEEE Transactions on Systems, Man and Cybernetics 1992; 22(3), (1992), 418-435.