Using a Rank Fusion Technique to Improve Shot Boundary Detection Effectiveness

M. Eduardo Ares and Álvaro Barreiro

IRLab, Department of Computer Science, University of A Coruña,
Campus de Elviña s/n, 15071, A Coruña, Spain
{maresb,barreiro}@udc.es
http://www.dc.fi.udc.es/irlab

Abstract. Achieving high effectiveness in Shot Boundary Detection (SBD) paves the way for high-level analysis of the video (keyframe extraction, story segmentation, etc.), which makes this step very important. Thus, the SBD problem has been extensively addressed, and many approaches have been proposed. As these approaches have their own different strengths and weaknesses, merging the outcomes of different detectors in order to obtain a better detector comes naturally. In this paper we propose an approach to SBD which takes into account the outcomes of two shot boundary detectors, using a rank fusion technique. This new detector is tested with videos from the TRECVid initiative, finding that it outperforms the two original methods. Moreover, the computation of this merging method is very fast, so it is very attractive for an operational environment.

1 Introduction

Nowadays, multimedia (and specially video) information has achieved a great importance in our society. Arguably, the best example of this situation is television, which has risen to a privileged position among the mass media, rivalling in followers and influence with newspapers. Besides, in the last few years, a great number of video repositories have appeared. Those popular webpages host an ever-growing collection of videos, letting users upload, search and browse them.

Consequently, in the last two decades there have been great efforts to develop techniques to perform automated video analysis and retrieval, like those available for text. Maybe the most significant proof of this trend was the concern that TREC [1] (Text Retrieval Conference) showed in that subject. That interest led in the first place to the creation of the TREC Video Track which in 2003 was turned into a initiative of its own, the TREC Video Retrieval Evaluation (TRECVid)[2], being the showcase of the new developments and advancements on automated video processing.

Almost all techniques developed for video analysis take the *shot* as retrieval unit. A shot is a video sequence which has been taken with the same camera without cuts in between, and is widely regarded as the minimal meaningful

R. Moreno-Díaz et al. (Eds.): EUROCAST 2009, LNCS 5717, pp. 247–254, 2009.

portion of a video, hence its importance to video analysis. Even when shots are not the outcome of the system (for instance, in TV News retrieval systems, where the retrieval unit is the *story*), the shots are used as building blocks of those bigger retrieval units. Thus, segmenting a video stream into shots is a very important process, whose effectiveness has a huge impact on the performance of the whole system. This process is called Shot Boundary Detection (SBD).

The boundary between shots can be categorised in two types, attending to the abruptness of the the transition: *hard cuts*, in which the transition happens instantly, and *gradual transitions*, in which the transition spans some frames. The detectors use the expected similarity between frames of the same shot, characterising the frontiers between shots as frames that bear a low visual similarity with previous ones. This similarity between frames is measured with features related with their visual content, being the main difference among SBD methods the choice of these features and the way they are used. Furthermore, the techniques used to detect one type of transition may (and often will) not be useful to detect the other type, which makes SBD even harder.

Due to its importance and difficulty, the SBD problem has been extensively addressed. The existing works have proposed a lot of features to measure the similarity of the frames: the plain absolute difference between pixels, the similarity between colour histograms[3], motion-compensated pixel differences[4], similarity between edge images[5]... The researches have also proposed many ways of using these features, which range from the simplest approaches, based on thresholding one feature, to more complex approaches, such as adaptive thresholding[3,6], statistical modelling[4], combining several features using rules[7,8] and machine learning techniques [9]. This intense research in the field and the great results that have been achieved have led to some authors to deem the SBD problem as almost resolved[10]. However, and precisely due to the problem's importance to the whole video retrieval process, there are still some aspects worth studying.

In this paper we propose an approach to SBD based on using a voting technique (the *Borda count*, which has been successfully used in metasearch to merge ranks[11]) to merge the outcomes of different detectors. Since each Shot Boundary Detector has its own strengths and weaknesses, creating a detector that improves their performance merging them comes naturally, making possible to reach higher effectiveness avoiding the need for a laborious parameter tuning. And even when the base performance is very good, a little improvement, whatever small it may be, is important. Furthermore, the merging method proposed is very fast and does not need any parameter tuning. Its simplicity and speed makes it very attractive for operational environments.

This work is focused on testing the feasibility of this method trying to merge the results of two simple shot boundary detectors to detect hard cuts, due to their relative simplicity compared with gradual transitions. The experimentation carried out showed that the aggregated method outperforms the methods it aggregates, providing an improvement which ranges from 3.5% to 15.6%.

Next, section 2 details the approach we are proposing; section 3 specifies how the evaluation was carried out; section 4 lists the results obtained in that evaluation and section 5 concludes and discusses the future work.

2 Details of the Approach

The approach to SBD proposed in this paper merges the outcomes of two detectors (a block matching detector and an edge-based detector) using a ranking merging method (the Borda count). These detectors were chosen because they are two well-known detectors, which are relatively simple and whose strengths are complementary.

2.1 Detector 1: Block Matching Detector

This detector uses a motion compensation approach similar to the one proposed in [4]. The next steps are followed when trying to compare frames f_i and f_{i+1}:

1. The frames' size is reduced (or reduced versions are taken directly from the video stream, using the DC coefficients).
2. The frames are splitted in n non-overlapping blocks
3. The best match for each block b_j of f_{i+1} is searched in f_i:
   - The search area for the best match is the region of f_i composed by the block of f_i which is located in the same position as b_j and the pixels which lay in a certain neighbourhood of that block.
   - The difference between two blocks b and b' is measured in terms of the square differences between their pixels in all colour channels

     $$\text{block difference}(b, b') = \sum_{x,y,c \in \text{channels}} (b(x,y,c) - b'(x,y,c))^2 \qquad (1)$$

   - For each b_j, the best matching block is the candidate which minimises this difference value.
4. The difference value between f_i and f_{i+1} is the sum of the difference values (1) between each b_j and its best matching candidate.

2.2 Detector 2: Edge-Based Detector

This detector is an implementation of one of the detectors proposed in [5]. It uses an approach based on comparing the edges detected in the frames. In order to compare two frames f_i and f_{i+1} the next steps are followed:

1. Both images are resized to half their size and converted to greyscale.
2. A Sobel edge detector is applied to both frames.
3. The edge images e_i and e_{i+1} are created from the results of step 2, taking the points whose edge intensity is greater than a certain threshold t as 1 (*edge*) and the others as 0 (*non-edge*).
4. A dilation morphological operation is applied to e_i and e_{i+1} ($\overline{e}_i$ and $\overline{e}_{i+1}$).

5. The ratios of "entering" (edge pixels in f_{i+1} which are not so in f_i) and "exiting" (edge pixels in f_i which are not so in f_{i+1}) edge pixels are calculated:

$$p_{in} = \frac{\sum_{x,y} e_{i+1}(x,y)\overline{e}_i(x,y)}{\sum_{x,y} e_{i+1}(x,y)}; p_{out} = \frac{\sum_{x,y} e_i(x,y)\overline{e}_{i+1}(x,y)}{\sum_{x,y} e_i(x,y)} \tag{2}$$

6. The difference value between f_i and f_{i+1} is the maximum of p_{in} and p_{out}.

In both detectors, the steps are applied all along the video to every pair of consecutive frames.

2.3 Rank Aggregation: Borda Count

The Borda count is a voting technique which has been successfully used in metasearch to merge document rankings [11]. We will use the simplest version of Borda count, where all voters are equal and their opinions are given the same weight. Given a set of n candidates and k voters, it follows the next steps:

1. each voter creates a ranking of the n candidates
2. each voter assigns n votes to the first candidate in its ranking, $n-1$ votes to the second, and so on.
3. the votes for each candidate are added
4. the aggregated ranking is created according to the scores calculated in 3.

The system we propose in this paper uses the Borda count in order to merge the outcomes of the two presented shot boundary detectors. Given a video composed by n frames (from f_1 to f_n), the *candidates* are the $n-1$ possible pairs composed by consecutive frames. Thus, the candidate c_1 would be composed by f_1 and f_2, c_2 would be composed by f_2 and f_3 and so on until c_{n-1}, which would be composed by f_{n-1} and f_n. On the other hand, the *voters* are the shot boundary detectors. For each candidate, the detectors calculate the difference value between the frames the candidate is composed of. Then, each detector ranks the pairs of frames according to this difference value, and each candidate is given a number of votes depending of its position in this ranking, where candidates with higher difference values are ranked higher. These votes are the only output of each detector. In other words, the difference values used by each detector are not considered further, avoiding the need for the normalisation of these scores (this is one of the most important advantages of Borda count). Once the votes of the two detectors are calculated, and as it was explained above, they are totalled. These sums are the outcome of the system, and the new difference values.

2.4 Cut Detection

The methods presented output a list of difference values for the transitions between each pair of frames. Once these values are calculated, a criterion has to be defined on how to use them to detect the cuts. A lot of criteria have been proposed, ranging from simple thresholding to adaptive approaches.

3 Evaluation

To test the feasibility of an approach to SBD based on ranking aggregation we will compare the effectiveness of the method explained in Sect. 2 with the two detectors (the Motion compensation based detector and the Edge based detector) on their own. As we are proposing and testing the first sketches of a new approach to combining evidences in SBD we have focused solely on the hard cuts, because of their simplicity compared with gradual transitions. Moreover, the hard cuts are the most common transition type in almost all video genres, so a good hard cut detection is mandatory for a successful SBD system. A proof of their importance are the ratios of transition types in the videos used in the SBD task of TRECVid. Since its inception in 2003 (and also in the TREC track) the hard cuts were the most common transition type, reaching in 2007 (the last year this task was considered) a ratio of 89.5%[10]. In that edition, two of the fifteen participating groups (U. Sheffield and U. Brno) focused exclusively in hard cuts.

In order to assess the effectiveness of our approach we have used a collection of five videos used in the TRECVid in years 2001 and 2005, which are in the public domain (Table 1). The human annotated ground truth marking the hard cuts present in these videos was obtained from the TRECVid page.

To measure the effectiveness of the detectors we have used precision (ratio of cuts which have been correctly detected) and recall (portion of the proper cuts on the video which have been detected), which are used in the TRECVid SBD task[10]. So as to summarise these two metrics we have used the F-measure (3), which gives the same importance to precision and recall.

$$\text{F-Measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

The detectors presented in Sects. 2.1 and 2.2 depend on some parameters. After testing several different settings, we chose the ones which showed the best results. For the block matching detector the reduction ratio of the frames was set to $\frac{1}{8}$ of the original size. The block size was set to 4×4 pixels, and the zone to search was 12 pixels. On the other hand, for the edge based detector the threshold of the intensity of the edges to detect the edges (t) was set to 200, and the structuring element used in the dilation step was a square of 3×3 pixels.

In the experiments the detectors explained in Subsects. 2.1, 2.2 and 2.3 were applied to the video collection. In the aggregated method a transition between a

Table 1. Video collection

Identifier	Video name	Length (frames)	Hard cuts
anni005	NASA 25th Anniversary Show, Segment 5	11364	38
anni009	NASA 25th Anniversary Show, Segment 9	12307	38
NASAConnect-HT	NASA Connect - Hidden	50823	143
NASAConnect-AO	NASA Connect - Ancient Observatories	51290	67
NASADT18	NASA Destination Tomorrow 18	51299	135

pair of frames was labelled as *cut* if and only if its number of votes was greater than a certain threshold t, which was the same for the whole video. In the other approaches the thresholding was made on the ranking position of the difference value of the pair of frames. For these methods, a threshold of t means that a transition was labelled as *cut* if and only if its difference value was above t difference values or more corresponding to other transitions of the video. Note that this has the same effect as thresholding the number of votes which that candidate would be given by that method in the aggregated approach. We have chosen this technique due to its homogeneity and simplicity, avoiding external factors which could influence the effectiveness of the system and the tuning of a more complex method. The best threshold for each method was set manually, testing a set of thresholds. For the aggregated method the thresholds tested ranged from 99.90% to 99.00% of the votes of the most voted candidate, while in the block matching and the edge based approaches the thresholds tested ranged from 99.90% to 99.00% of the total number of transitions. Then, the threshold which reached the best F-Measure was selected. This methodology was chosen in order to test the effectiveness of the proposed rank aggregation technique, regardless of the threshold calculating method. Obviously, this approach is not suitable for real world applications, where the threshold must be calculated automatically.

4 Results

We present in Table 2 the results of the three methods tested (block matching, edge based and aggregated) for each video in the collection. The results shown are the F-Measures obtained with the best threshold (see Table 3) and the improvement achieved by the aggregated method over the best performing of the two plain shot boundary detectors.

4.1 Discussion

In all videos the aggregated method we propose in this paper outperforms the methods it is composed of. Even though this was the expected behaviour, there are some aspects worth remarking:

First, and even though the effectiveness of the aggregated method is obviously dependant on the effectiveness of the methods merged, it provides an improvement of the performance which ranges from 3.4% to 15.6%. This improvement

Table 2. F-measure for each method and improvement achieved over the best performing detector

Video	anni005	anni009	NasaConnect-HT	NasaConnect-AO	NASADT18
Block matching	0.904	0.880	0.818	0.773	0.692
Edge	0.932	0.640	0.761	0.255	0.591
Borda count	0.961	0.911	0.930	0.887	0.800
Improvement	3.1%	3.5%	5.0%	11.4%	15.6%

Table 3. Cut ratios and Best thresholds for all videos

Video	Cuts ratio	Best threshold
anni005	00.33%	99.60%
anni009	00.31%	99.30%
NasaConnect-HT	00.29%	99.60%
NasaConnect-AO	00.13%	99.60%
NASADT18	00.26%	99.60%

is higher in the videos where the two simple methods perform worse. It should be also noted that this improvement of the results is achieved in all videos, regardless of the difference of effectiveness between the proposed detectors. In the results of NasaConnect-AO it can be noted how the outcomes of a detector which performs fairly well and another one which performs very poorly are merged without degrading the performance of the first one, and in fact improving it. This shows the robustness of the merging method chosen.

Moreover, as shown in Table 3, the best threshold for the aggregated method is stable along the collection (almost always the 99.60% of the maximum value), regardless of the varying cut ratio. This fact should be taken into account when devising a strategy to calculate automatically the threshold.

In order to assess the statistical significance of the results we have performed a Wilcoxon test. The hypothesis were H_0: our method does not outperform the individual detectors and H_1: our method is better than the individual detectors. The test showed that the aggregated method is significantly better than the best of the others detectors in each video with a p-value < 0.05.

5 Conclusions and Future Work

In a general level, we should note that, as it has been previously stated, the objective of this work was to test the feasibility of a method to merge the outcomes of different shot boundary detectors based on the Borda count. According to the results presented in Sect. 4, its feasibility is proved. It is also worth remarking the great improvement achieved by a method which merges blindly the outcomes of other detectors, despite not having previous training or being imbued with domain knowledge. It should be also noted the simplicity and the low computational cost of the merging method: once the difference values are calculated, the time invested in merging them is negligible. Also, we are avoiding the need for a normalisation of the difference values and the problems that would be associated with that process, such as modelling of the outcomes of each method.

Future work should be aimed in two main directions: testing the behaviour of this method when trying to detect gradual transitions and developing a method to set automatically the threshold of the detection. Also, the method should be tested against a bigger video collection and with more detectors.

Moreover, and centring in the method itself, there are two main questions still worth addressing. The first one is trying to devise a way to avoid having to wait until the whole video is processed to have results. This could be a problem if there are restrictions of time (i.e. a cut should be detected in no more than a certain time since it happens) and specially if we are trying to segment long videos or if the detectors we are trying to merge are computationally expensive.

The second aspect that we think worth studying is using a weighted Borda count approach, which would enable us to weight more the votes of those detectors which seem more reliable.

Acknowledgements. This work is co-funded by *Ministerio de Ciencia e Innovación*, FEDER and *Xunta de Galicia* under projects TIN2008-06566-C04-04 and 07SIN005206PR. M. Eduardo Ares also wants to acknowledge the support of the FPU programme of *Ministerio de Educación* of Spanish Government.

References

1. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing). MIT Press, Cambridge (2005)
2. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: Proceedings of MIR 2006, pp. 321–330 (2006)
3. O'Toole, C., Smeaton, A.F., Murphy, N., Marlow., S.: Evaluation of automatic shot boundary detection on a large video test suite. In: Proceedings of CIR 1999 (1999)
4. Hanjalic, A.: Shot-boundary detection: unraveled and resolved? IEEE TCSV 12(2), 90–105 (2002)
5. Smeaton, A.F., Gilvarry, J., Gormley, G., Tobin, B., Marlow, S., Murphy., N.: An evaluation of alternative techniques for automatic detection of shot boundaries in digital video. In: Proceedings of IMVIP 1999 (1999)
6. Yeo, B.L., Liu, B.: Rapid scene analysis on compressed video. IEEE TCSV 5(6), 533–544 (1995)
7. Browne, P., Smeaton, A.F., Murphy, N., O'Connor, N., Marlow, S., Berrut, C.: Evaluating and combining digital video shot boundary detection algorithms. In: Proceedings of IMVIP 2000, pp. 93–100 (2000)
8. Liu, Z., Gibbon, D., Zavesky, E., Shahraray, B., Haffner, P.: A fast, comprehensive shot boundary determination system. In: Proceedings of IEEE ICME 2007, pp. 1487–1490 (2007)
9. Matsumoto, K., Naito, M., Hoashi, K., Sugaya, F.: SVM-based shot boundary detection with a novel feature. In: Proceedings of IEEE ICME 2006, pp. 1837–1840 (2006)
10. Over, P., Awad, G., Kraaij, W., Smeaton., A.F.: TRECVID 2007 overview. In: TRECVid 2007 - Text REtrieval Conference TRECVid Workshop (2007)
11. Aslam, J.A., Montague, M.: Models for metasearch. In: Proceedings of SIGIR 2001, pp. 276–284 (2001)