RMIT University Video Shot Boundary Detection at TRECVID 2005

Timo Volkmer S.M.M. Tahaghoghi

School of Computer Science and Information Technology
RMIT University, GPO Box 2476V, Melbourne, Australia, 3001

{tvolkmer,saied}@cs.rmit.edu.au

Run overview

We participated in the Shot Boundary Detection task. This page provides a summary of: (1) the approaches tested in the submitted runs; (2) differences in results between the runs; (3) the overall relative contribution of the techniques; and, (4) our overall conclusions.

1. Our approach to shot boundary detection uses the moving query window technique [8, 9, 10]. In 2005, we have applied a new implementation of our system and experimented with different feature representations. We submitted ten runs using only visual features, exploring different colour histogram representations. The first two runs were used as a baseline in which we have used our system as it was applied in 2004, with the settings as in our best runs of that year (Run 3 and Run 5) [11]. An overview of all submitted runs is shown in Table 1.

	Feature		HWS		Threshold
Run	cuts	gradual	cuts	gradual	method
1	HSVl	HSVl	6	14	old
2	HSVl	HSVl	8	16	old
3	HSVl	HSVl	8	14	new
4	HSVl	HSVl	10	16	new
5	HSV3	HSVl	8	14	new
6	HSV3	HSVl	10	16	new
7	HSV3	HSV3	8	14	old
8	HSV3	HSV3	10	16	old
9	HSVl	HSV3	8	14	old
10	HSVl	HSV3	10	16	old

*Table 1: Overview over our ten submitted runs in 2005, the features that we have used, and variations in the settings for the half-window size (*HWS*). Runs 1 and 2 were carried out with our 2004 system and serve as a baseline.*

2. In our submissions we have tested a new implementation of our system that is designed as a two-pass algorithm, rather than the single-pass algorithm used in previous years. We have applied different combinations of a localised HSV histogram (HSVl) feature and a true three-dimensional colour histogram (HSV3) representation in Run 3 through to Run 10. We have also implemented a new dynamic threshold computation that was applied in Run 3 through to Run 6. This comes into effect during gradual transition detection and is designed to minimise the number of false positives in clips with few transitions.

3. Our three-dimensional colour histogram expresses each colour as a point in the three-dimensional space. While this representation has been shown to produce promising results in content-based image retrieval, performance gains are often outweighed by computational overhead. Due to the type of footage in 2005, the new threshold computation has had only very limited influence on the results.

4. The baseline runs which performed very well in 2004 were again our best runs. Despite improved results during training on the 2003 test set with our new implementation, we could not achieve improvements on the 2005 test set. We see the reasons for this mainly in the limited training that we were able to undertake with our new two-pass algorithm and the different feature combinations.

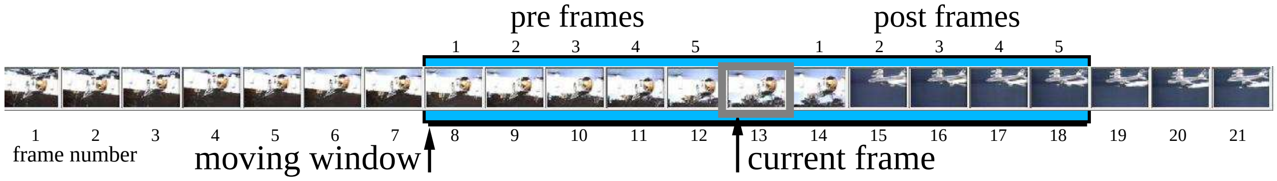

*Figure 1: Moving query window with a half-window size (*HWS*) of 5. The five frames before and the five frames after the current frame form a collection on which the current frame is used as a query example.*

1 Introduction

The task of identifying the basic semantic entities in video clips — the *shots* [4] — is crucial for enabling search and retrieval. It is usually achieved through finding the transitions that constitute the boundaries between adjacent shots. Most automatic shot boundary detection methods make use of the observation that frames are similar when they are within a shot, and dissimilar when they span a transition. This generally holds for all kinds of transitions, regardless of whether they are abrupt or gradual.

The accurate detection of gradual transitions still constitutes a more challenging problem [5, 6] compared to the detection of abrupt transitions, that is, cuts [1, 7, 9]. Results of standardised benchmarks such as the TRECVID shot boundary detection task support this observation. In this paper, we present our technique for shot boundary detection. Using a new implementation of our moving query window from previous TRECVID workshops, we have experimented with different one-dimensional and three-dimensional histogram representations for frame comparison. We report and discuss results obtained when we applied our system in the TRECVID 2005 shot boundary detection task.

2 Our approach

We use our moving query window approach previously presented at TRECVID [8, 11, 12]. However, this year we have used a new implementation of this method and experimented with different histogram representations. Figure 1 shows a moving query window: the window has an equal number of frames on either side of a current frame, and the current frame is advanced or *moved* as the video is processed.

Each frame is represented by a colour histogram that is extracted from the video stream in a preprocessing step. We used two types of histogram: one, localised HSV histograms with 16 equal-sized regions; and two, three-dimensional global HSV histograms, where each colour is represented as a point in a three-dimensional space. For the localised histograms, we evaluate the quantised pixel information for all colour components within a single vector dimension by applying the Manhattan distance measure [2].

The information in the three-dimensional histograms is evaluated as a three-dimensional vector. Here, we compute the Manhattan distance over all three dimensions.

We believe that detecting gradual transitions is rather different to detecting abrupt transitions, and so our implementation allows us to apply a different evaluation process for each type of transition. Cut detection is performed in the first pass of our two-pass algorithm; gradual transitions are detected during the second pass. We continue with a discussion of the details of our approach.

2.1 Abrupt transitions

For cut detection, we use our ranking-based method [9]. This has been proven to work very effectively [8] with features derived from the Daubechies wavelet transform [3]; however, computation of wavelets is expensive. In 2003, to reduce computational cost, we used the ranking-based method in combination with one-dimensional, global histograms in the HSV colour space [12]. Results were strong, but not as good as those obtained with the wavelet feature. We used localised HSV histograms with 4×4 frame regions in 2004. By disregarding the frame centre during cut detection, we were able to achieve significant improvements over 2003 [11].

In our experiments in 2005 we have kept the same technique for cut detection. We have additionally experimented with global, three-dimensional HSV histograms that do not allow us to apply any localisation within one frame. In contrast to our strategy when using localised histograms, we do not disregard any part of the frame, and compute the inter-frame difference based on the entire feature vector of one frame.

2.2 Gradual transitions

Gradual transition detection in our system is based on comparing average frame similarity within the moving query window [10, 11]. In contrast to 2004, our new implementation is designed as a two-pass algorithm. This allows us to perform cut detection in an initial pass, during which we also calculate a noise factor over the entire video. During gradual transition detection in the second pass, we use this noise factor to adjust the dynamically computed threshold that we apply for peak detection.

Some of the videos in our training set, the TRECVID 2003 shot boundary test set, were characterised by low activity and very few or no gradual transitions. Our previous system had a tendency to produce many false positive detections in such cases. The noise-factor based threshold adjustment has shown significant improvements of our results during training.

As for cuts, we have experimented with different histogram representations for gradual transition detection this year. First feature was again the one-dimensional, localised HSV histogram, in which we divide each frame into 16 equal-sized regions. For gradual transitions, we use identical weights for each frame region, that is, we do not disregard any part of a frame. The distance between frames is computed based on the average distance of all corresponding regions.

A global, three-dimensional colour histogram in the HSV colour space is the second feature, in which each colour is represented as a point in a three-dimensional space. We compute the inter-frame difference by applying an unweighted Manhattan distance measure over all three dimensions of the feature vectors.

2.3 Algorithm details

Most parameters are either dynamically computed during processing or pre-determined experimentally. Those parameters that we have determined on previous training results include, for example, the Threshold History Size (THS) [11], which we have set to the size of the query window. We have fixed the Upper Threshold Factor (UTF) factor to 1.7 [11] for all runs in which we do not use the new threshold computation method. Our new thresholding method does not require any preset parameters.

An important remaining parameter of our system is the size of the moving window: we refer to this as the Half-Window Size (HWS), that is, the number of frames on either side of the current frame. We have experimented with different sizes for cut detection and gradual transition detection. We have previously determined optimum settings for television news footage [9, 11] but for a test set with varying types, as in 2005, it is difficult to determine one optimal window size. We have therefore varied the half-window size in our experiments between 6 and 10 for cut detection, and 14 and 16 for gradual transition detection. The settings we used are detailed in Table 1.

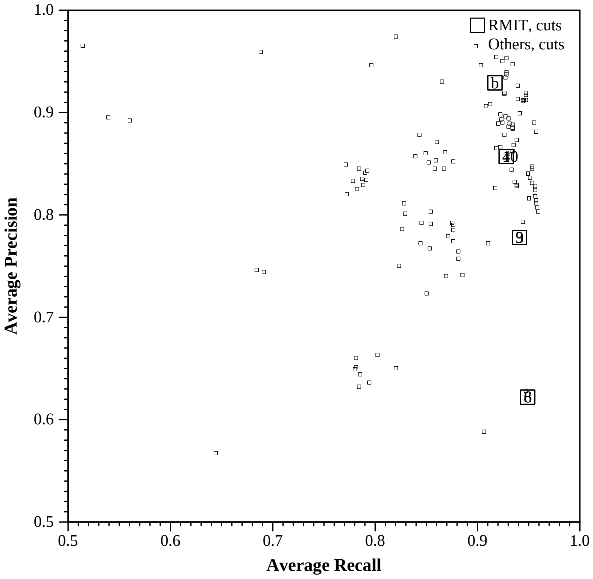

Figure 2: Performance of our system for abrupt transitions on the TRECVID 2005 *shot boundary detection task, measured by Recall and Precision.*

3 Results and Discussion

In this section, we discuss results of our system for shot boundary detection when applied to the TRECVID 2005 test set.

The test set for shot boundary detection consisted of 12 video files with a total duration of approximately 6 hours. There were 4,535 transitions, of which 2,759 were cuts and 1,776 were gradual transitions. The test collection is comprised of American, Chinese, and Arabic television news footage, interrupted by commercials and entertainment segments. Four video clips are promotional films by NASA.

Figure 2 shows the performance of our system for cut detection, measured in recall and precision, and compared to all other submissions. Interestingly, the results are only competitive for the baseline runs.

The combined results in cut and gradual transition detection — as shown in Figure 3 — show that overall, our system is not as competitive as last year. Our

	HWS		All Transitions		Cuts		Gradual Transitions			
Run	cuts	grad.	Recall	Precision	Recall	Precision	Recall	Precision	F-Recall	F-Precision
b	6	14	0.891	0.818	0.917	0.929	0.816	0.588	0.753	0.804
3	8	14	0.884	0.730	0.941	0.778	0.719	0.592	0.777	0.810
4	10	16	0.879	0.794	0.928	0.857	0.733	0.624	0.815	0.804
5	8	14	0.895	0.511	0.954	0.490	0.721	0.609	0.773	0.810
6	10	16	0.894	0.627	0.949	0.622	0.732	0.645	0.811	0.805
7	8	14	0.922	0.503	0.954	0.490	0.830	0.548	0.708	0.806
8	10	16	0.920	0.603	0.949	0.622	0.835	0.548	0.746	0.801
9	8	14	0.912	0.702	0.941	0.778	0.829	0.531	0.711	0.806
10	10	16	0.905	0.749	0.928	0.857	0.836	0.530	0.750	0.802

Table 2: Detailed results for all runs for various settings of the half-window size HWS. *Results for the two baseline results were nearly identical and are combined in the first row.*

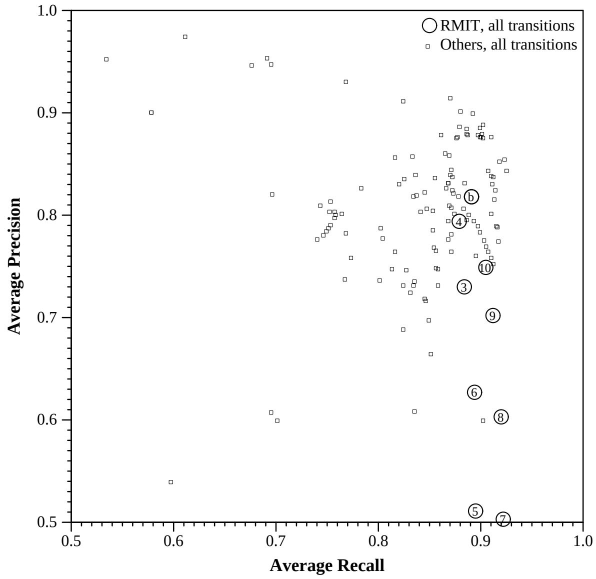

Figure 3: Performance of our system for all transitions on the TRECVID 2005 *shot boundary detection task, measured by Recall and Precision.*

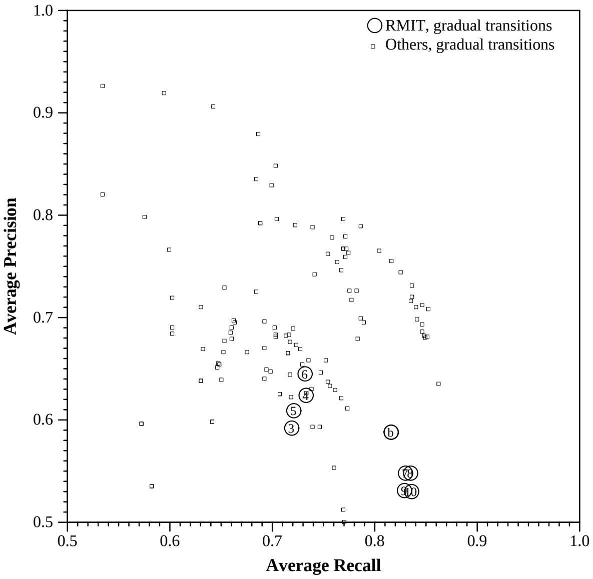

Figure 4: Recall and Precision of our system for gradual transition detecion on the TRECVID 2005 *shot boundary detection task.*

best runs were the baseline runs produced by our 2004 configuration.

We were unable to achieve an improvement through the new two-pass algorithm. Indeed, the previous single-pass algorithm tends to produce fewer false positives, which aids precision. In particular, our system produced many false positives on one of the NASA promotional video clips that had very few cuts. We have yet to exactly determine what causes this effect; at this point, we cannot entirely exclude a problem with the implementation.

In Figure 4, the recall and precision of our technique for gradual transitions is shown compared to the results of the other submissions. Results are not as strong as last year mainly because of a weak precision in all runs. Our previous implementation performed both cut detection and gradual transition detection during a single pass, with a bias towards detection of cuts to avoid false detection of gradual transitions. The new implementation does not have this bias, and the consequent loss in precision is not offset by improvements offered by the new thresholding algorithm.

The three-dimensional histogram feature does not seem to be robust across different types of footage, and all runs that use it appear towards the lower end of our scale.

Figure 5 shows frame recall and frame precision to measure how accurately we detect the start and end of gradual transitions. We observe good results, similar to last year. Table 2 shows detailed results of all our

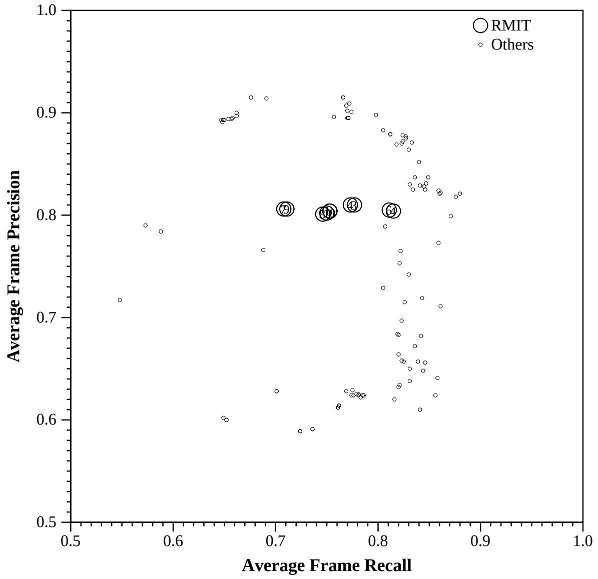

Figure 5: Frame Recall and Frame Precision of our system for gradual transitions on the TRECVID 2005 *shot boundary detection task.*

Run	Total	Decoding	Segmentation
1	22,391.5	22,135.2	296.7
2	22,446.7	22,135.2	311.5
3	13,233.4	12,932.5	300.8
4	13,284.2	12,932.5	351.8
5	9,729.7	9,137.6	592.0
6	9,867.7	9,137.6	730.2
7	10,239.2	9,137.6	1,101.4
8	10,431.3	9,137.6	1,293.7
9	13,784.8	12,932.5	852.3
10	13,918.9	12,932.5	986.7

Table 3: Timing results in seconds (real time) for all runs of our shot boundary segmentation system.

runs along with the parameter settings that we have used for the Half-Window Size (HWS).

Timing results are shown in Table 3. To exclude the time that the system spends on task switching and other processes, we calculated elapsed time in seconds based on the sum of user time and system time when processing our runs. The timing experiments were performed on a single CPU machine with an Athlon-64 3200+ (2.2 GHz) processor, 1,024 MB of main memory, and running SuSE Linux 9.3 with the SuSE standard kernel 2.6.11.

The bulk of the processing time is clearly spent on decoding the video stream. Our segmentation algorithm can be implemented very efficiently. Our current implementations, however, are not optimised for efficiency.

4 Conclusion and Future Work

We have presented our approach to shot boundary detection in form of a new two-pass implementation and experimented with different histogram representations. The fact that our baseline runs — carried out with our 2004 system — performed best shows that the approach as such works well. Applying this to a new test set without any improvements and with only limited training, we have achieved competitive results in both cut detection and gradual transition detection. However, the question remains why we could not transfer the good training results with our new implementation to the 2005 test set. We believe one reason is the two-pass design; we intend to explore whether it would be appropriate to revert to a single-pass system. Another reason is our choice of features; the global three-dimensional histogram representation does not appear to be as robust across different kinds of footage. Interestingly, our cut detection stage in the new algorithm failed for one particular video; we plan to address this in future work.

References

[1] B. Adams, A. Amir, C. Dorai, S. Ghosal, G. Iyengar, A. Jaimes, C. Lang, C.-Y. Lin, A. Natsev, M. Naphade, C. Neti, H. J. Nock, H. H. Permuter, R. Singh, J. R. Smith, S. Srinivasan, B. L. Tseng, T. V. Ashwin, and D. Zhang. IBM Research TREC-2002 video retrieval system. In E. M. Voorhees and L. P. Buckland, editors, *NIST Special Publication 500-251: Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, pages 289–298, Gaithersburg, MD, USA, 19–22 November 2002.

[2] D. Androutsos, K. N. Plataniotis, and A. N. Venetsanopoulos. Distance measures for color image retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'98)*, volume 2, pages 770–774, Chicago, IL, USA, 4–7 October 1998.

[3] I. Daubechies. *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[4] A. Del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.

[5] R. W. Lienhart. Reliable transition detection in videos: A survey and practitioner's guide. *International Journal of Image and Graphics (IJIG)*, 1(3):469–486, July 2001.

[6] S. Marchand-Maillet. Content-based video retrieval: An overview. Technical Report 00.06, CUI - University of Geneva, Geneva, Switzerland, 2000.

[7] G. M. Quénot, D. Moraru, and L. Besacier. CLIPS at TRECVID: Shot boundary detection and feature detection. In E. M. Voorhees and L. P. Buckland, editors, *TRECVID 2003 Workshop Notebook Papers*, pages 35–40, Gaithersburg, MD, USA, 17–18 November 2003.

[8] S. M. M. Tahaghoghi, J. A. Thom, and H. E. Williams. Shot boundary detection using the moving query window. In E. M. Voorhees and L. P. Buckland, editors, *NIST Special Publication 500-251: Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, pages 529–538, Gaithersburg, MD, USA, 19–22 November 2002.

[9] S. M. M. Tahaghoghi, J. A. Thom, H. E. Williams, and T. Volkmer. Video cut detection using frame windows. In V. Estivill-Castro, editor, *Proceedings of the Twenty-Eighth Australasian Computer Science Conference (ACSC 2005)*, volume 38, Newcastle, NSW, Australia, 31 January – 3 February 2005. Australian Computer Society. To appear.

[10] T. Volkmer, S. M. M. Tahaghoghi, and H. E. Williams. Gradual transition detection using average frame similarity. In S. Guler, A. G. Hauptmann, and A. Henrich, editors, *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop (CVPR-04)*, Washington, DC, USA, 2 July 2004. IEEE Computer Society.

[11] T. Volkmer, S. M. M. Tahaghoghi, and H. E. Williams. RMIT University at TRECVID-2004. In *TRECVID 2004 Workshop Notebook Papers*, Gaithersburg, MD, USA, 2004.

[12] T. Volkmer, S. M. M. Tahaghoghi, H. E. Williams, and J. A. Thom. The moving query window for shot boundary detection at TREC-12. In *TRECVID 2003 Workshop Notebook Papers*, pages 147–156, Gaithersburg, MD, USA, 17–18 November 2003.