University of Sheffield at TRECVID 2007: Shot Boundary Detection and Rushes Summarisation

Siripinyo Chantamunee Yoshihiko Gotoh
Department of Computer Science, University of Sheffield, United Kingdom
{*s.chantamunee, y.gotoh*}@*dcs.shef.ac.uk*

Abstract

This year we conducted experiments on shot boundary detection and rushes video summarisation. For the shot boundary determination task, we focused on detection of '*cut*'. The approach calculated the '*exclusive or*' of two frames in the grey scale in order to measure the amount of discontinuity at a pixel level between two shots. Five runs were submitted with different sets of parameters, resulting in the performance of as high as 87% recall and 85% precision. For the rushes video task, the summary duration was fixed at 4% of the video length. We joined a number of continuous frames that were extracted from the middle of each shot detected. We submitted a single run, resulting in the average level performance.

1 Shot Boundary Detection

Shot boundary detection is a process of identifying boundaries between shots from a sequence of video frames. The key idea is to choose a right set of features and measures that capture the dissimilarity between shots. The difference between adjacent frame pair is calculated from features. A shot boundary is assigned when the value is greater than a predefined thresholds.

To date there have been a large number of shot boundary detection algorithms proposed (Pye *et al.*, 1998; Lienhart, 2001; Pickering & Ruger, 2001; Deng & Manjunath, 2001; Miene *et al.*, 2003; Qi *et al.*, 2003; Quenot *et al.*, 2003; Ren & Singh, 2004; Ngo *et al.*, 2005). Features such as motion, colour, and edges have been tested using a number of different criteria for comparison. In general, the difficulty of the task depends on the complexity of shot transition, video structure and quality. We present the approach to detecting '*cut*', the most frequent shot transitions. It is relatively simple to identify '*cut*' because a clear discontinuity can be observed at a pixel level. Thus our approach assigns shot boundaries based on pixel changes between two adjacent shots.

1.1 Approach

Our approach aims to measure the amount of discontinuity at a pixel level between two consecutive frames. It is captured by calculating the '*exclusive or*' of two frames in the grey scale:

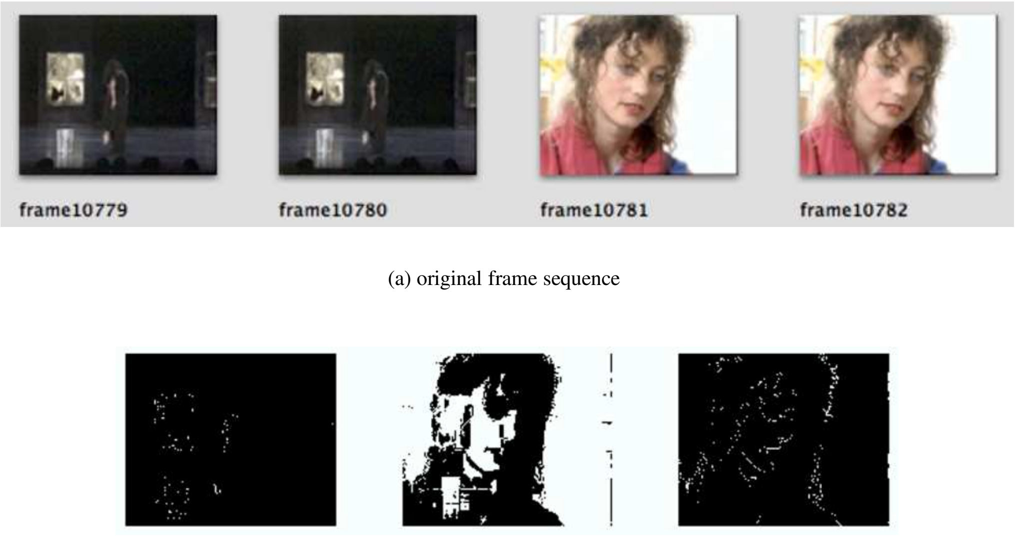
$$BWX_i = \sum_{j=1\dots J} P_{i-1}(j) \oplus P_i(j) \tag{1}$$

where $\oplus$ implies the '*exclusive or*' operation, J is the total number of pixels per frame, and $P_{i-1}(j)$ and $P_i(j)$ are the black/white value of pixel j in frames $i-1$ and i respectively. If the value for BWX_i is greater than a predefined threshold, '*cut*' is assigned between frames $i-1$ and i. A sample of this operation is illustrated in Figure 1. For the test video '*BG 34901*', Figure 2 demonstrates the calculation of BWX_i and $(BWX_{i+1} - BWX_i)$ for the entire length of video. In the experiment, each spike is treated as a shot boundary candidate.

1.2 Experiment

The Netherlands Institute for Sound and Vision provided a sound and vision collection including news magazine, science news, reports, documentaries, and educational programming (TRECVID, 2007). The test collection consisted of 15 mpeg-1 videos with the total length of six hours. The *visualDub*[1] software was applied to extract individual RGB frames of 352×288 pixels from the collection at a rate of 25 frames/second.

[1] available at http://www.virtualdub.org/



(a) original frame sequence

(b) ‘*exclusive or*’ of the black/white values between two adjacent frames

Figure 1: Detection of ‘*cut*’ using the ‘*exclusive or*’.

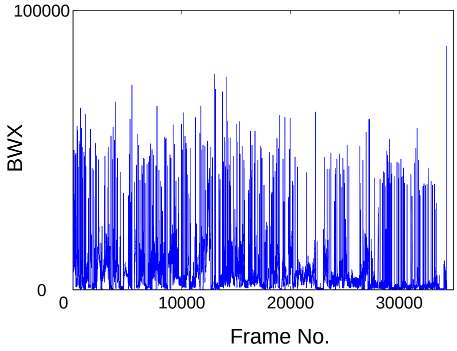


Figure 2: Test data ‘*BG 34901*’: calculation of BWX_i.

We submitted five runs. The first three runs were based on the calculation of BWX_i/J, and ‘*cut*’ was assigned for transitions:

UShefID1: BWX_i/J with the threshold value of 0.2;

UShefID2: BWX_i/J with the threshold value of 0.3;

UShefID3: BWX_i/J with the threshold value of 0.5.

For two more runs, a window of 100 frames was applied and moved forward by overlapping every 50 frames. For each of individual windows, $(BWX_{i+1} - BWX_i)$ were calculated, and ‘*cut*’ was assigned when

UShefID4: $(BWX_{i+1} - BWX_i) > \mu + 2\sigma$ where μ and σ are the mean and the standard deviation;

UShefID5: the maximum value of $(BWX_{i+1} - BWX_i)$.

1.3 Results and Discussion

Our submissions were evaluated by NIST. Table 1 shows ‘*cut*’ detection result showing the precision and recall. Among our five runs, the second run, using BWX_i/J with the threshold value of 0.3 (*i.e.*, a shot boundary was

	recall	precision
UShefID1	0.949	0.660
UShefID2	0.872	0.850
UShefID3	0.297	0.910
UShefID4	0.965	0.443
UShefID5	0.796	0.894

Table 1: ‘*Cut*’ detection results for our five runs, evaluated by NIST. Recall and precision values are the average for 15 test videos.

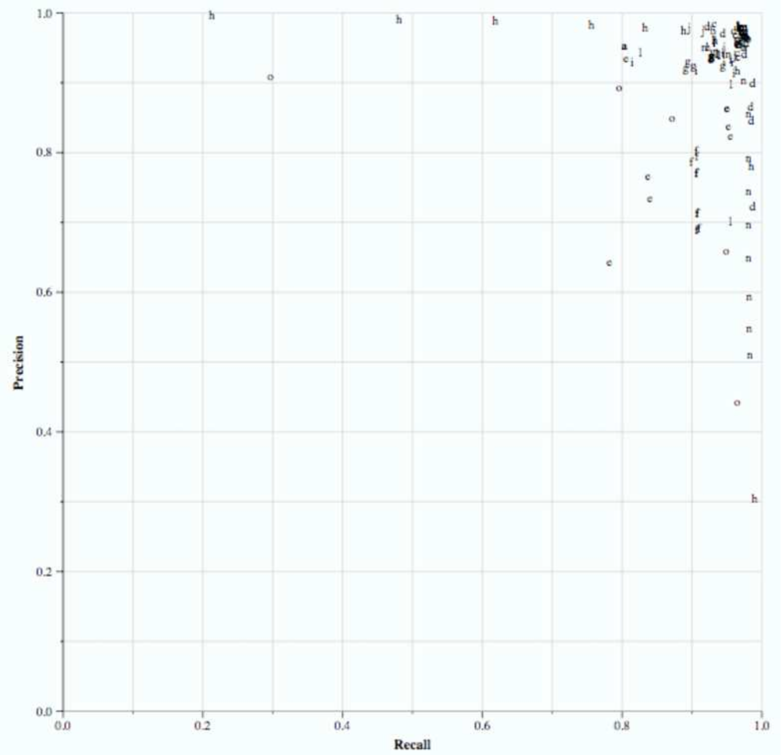


Figure 3: Precision and recall for ‘*cut*’ transition. Our five runs were labelled with ‘o’.

assigned when there were at least 30% of pixels different between two consecutive frames), performed best at 87% recall and 85% precision.

Figure 3, provided by NIST, shows our runs require further improvement in comparison to other approaches. None of our runs involved expensive calculation however, because we applied ‘*exclusive or*’ on the grey scale values, the method failed to identify a ‘*cut*’ when the background in both shots contain black or white pixels in the large area. It may be interesting to experiment with different colour spaces to address this problem. Use of keyframes may be another idea for improvement if they are available.

2 Rushes Summarisation

Rushes (or pre production video) is a raw material and is further processed to produce video data such as movies and television programmes. The material contains natural sound and highly repetitive frame sequences. Many retakes of the same scene duplicate the size of rushes to many times of the final video. Duplicated sequences are often caused by video production errors — *e.g.*, actors performing the incorrect line of a story and low quality audio visual contents (Smeaton *et al.*, 2006). The nature of rushes indicates that we require sophisticated technologies for managing and accessing the contents.

NIST launched the rushes summarisation task for this year’s TREC video evaluation (NIST, 2007). It aims to automate the creation of a summary clip from rushes video. It is required that a summary includes major objects and events specified by NIST. The objectives of evaluation are (1) to minimise the number of frames in a summary video and (2) to present information in a way that maximises the usability of the summary (TRECVID, 2007). We submitted a single run for a test data set, which was evaluated by NIST using seven measurements.

2.1 Approach

Our assumption is that a shot is the basic structure of video and that each shot contains important contents in the middle. A short frame sequence (or clip), not longer than 4% of the original shot, is extracted from the middle of each shot. Clips from multiple shots are then concatenated to form a summary. As a consequence, the summary length is 4% of the original rushes.

To determine the position of shot boundaries, a colour histogram based technique is employed. For each frame, the RGB colour histogram with 256 colour bins is extracted. The difference is calculated for each pair of two consecutive frames (Nagasaka & Tanaka, 1991):

$$DH_i = \sum_{k=1 \ldots K} |H_{i-1}(k) - H_i(k)| \tag{2}$$

where K is the number of bins, and $H_{i-1}(k)$ and $H_i(k)$ are counts for k^{th} bin in frames $i-1$ and i respectively. A shot boundary is assigned if DH_i is greater than a predefined threshold.

2.2 Experiment

The test collection consisted of 42 mpeg-1 rushes video, the majority of which were drama series and documentaries. The 352×288-pixel video frames were decoded in the rate of 25 frames/second. We completed a single run of the experiment using the shot reference based approach described above. To detect shot boundaries, a window of 100 frames was set in order to calculate the average value of all colour histogram difference values in the range. If the value was greater than 50% of the average, a shot boundary was assigned. The window was then moved forward with 50 frames overlap.

The evaluation was subjective, and performed by NIST with seven assessors. Each summary was judged by three assessors. Two baselines were provided by Carnegie Mellon University (CMU). Of which the uniform baseline was created by selecting one second of a frame sequence for every 25 seconds in the original rushes video. The second baseline summary, the colour cluster, was created from K-means clusters of fixed numbers based on HSV colour histogram, where the segment, closet to the centroid, was selected.

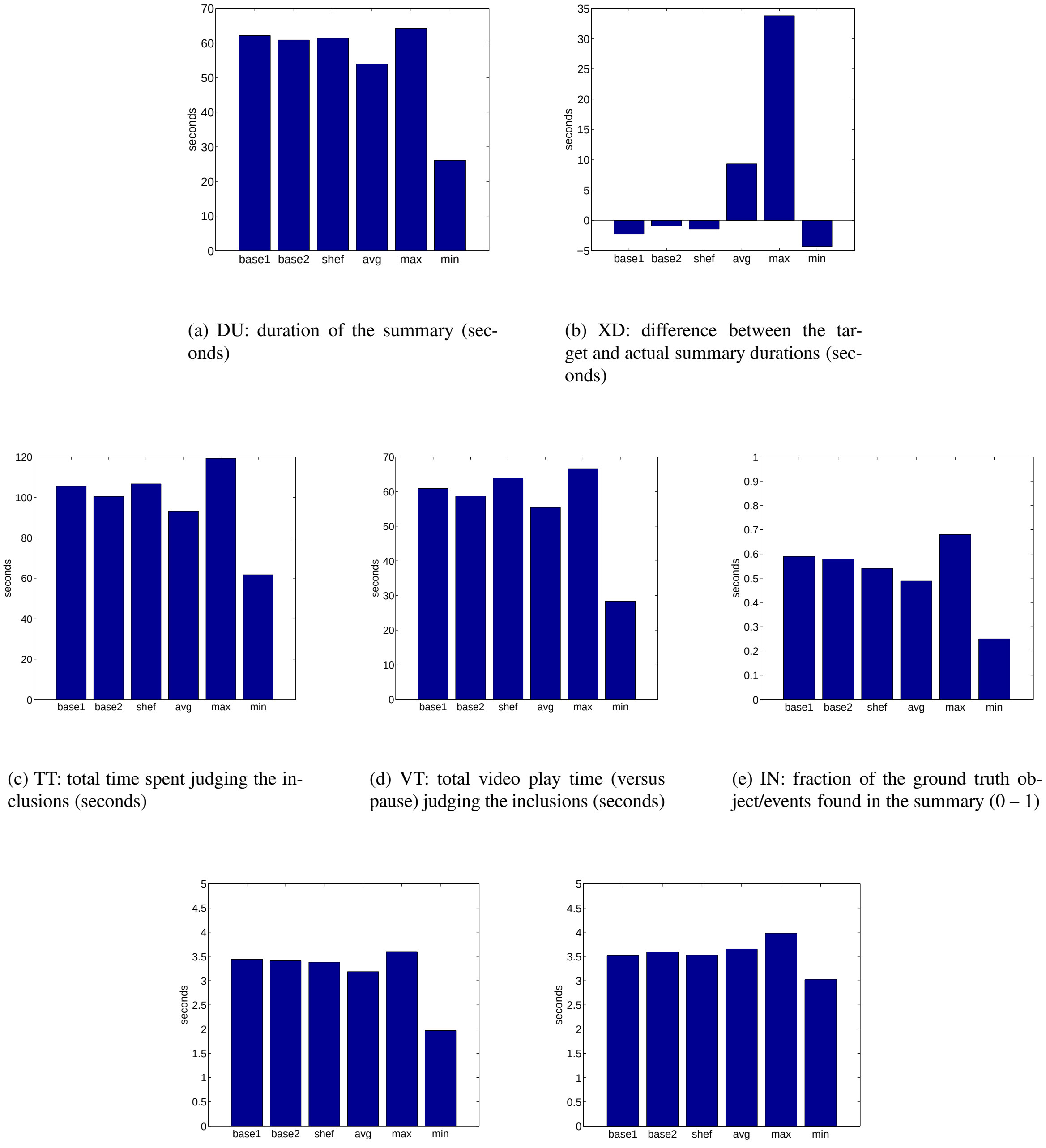
2.3 Results and Discussion

The test data set consisted of 42 rushes videos. Figure 4 compares two baselines, our single run, the average, the maximum, and the minimum of all runs submitted by various institutions. They were measured by seven criteria. Although the approach was simple, our run performed fair in comparison to other submissions. Our approach was similar to the CMU colour cluster baseline. It seems to imply that colour was simple but good feature to differentiate video contents. Use of different colour spaces (RGB for ours, HSV for CMU) did not influence on overall performance.

On average, 54% of the groundtruth were included in our run. Performance for individual summaries varied from 91% to below 30% (three summaries). Although inclusion of the groundtruth is an important factor for a summarisation task, other measures should also be taken into consideration to reflect the performance. The amount of duplicated contents, in particular, may be a crucial indicator for automatic summarisation of rushes video. Our run did not filter out duplicated contents, thus resulted in a weak score of 3.53. Because the approach extracted a short frame sequence in the middle of each shot, the summary contents were moderately understandable.

3 Conclusions

We participated in TRECVID 2007 with the shot boundary detection and the rushes video summarisation tasks. In shot boundary detection, we presented a pixel-based approach using the '*exclusive or*' of two frames in the grey scale. Our best result was 87% in recall and 85% in precision, which requires further improvement. For rushes video summarisation, a frame sequence was extracted from the middle of every detected shot. The summary clip was the concatenation of all frame sequences extracted from rushes video. It was found that rushes contained



(a) DU: duration of the summary (seconds)

(b) XD: difference between the target and actual summary durations (seconds)

(c) TT: total time spent judging the inclusions (seconds)

(d) VT: total video play time (versus pause) judging the inclusions (seconds)

(e) IN: fraction of the ground truth object/events found in the summary (0 – 1)

(f) EA: level of easy to understand (1 strongly disagree – 5 strongly agree)

(g) RE: level of duplication (1 strongly agree – 5 strongly disagree)

Figure 4: Evaluation result for the rushes summarisation task. Summaries were evaluated by NIST using seven criteria. Our single run (shef) is compared with the uniform baseline (base1), the colour cluster baseline (base2), the average (avg), the maximum (max), and the minimum (min) of all runs by various institutions.

many repetitive contents which strongly affected the quality of the summary. We expect the performance can be improved by filtering redundant shots, such as colour-bar, black/grey frames and repetitive shots.

References

Deng, Y. & Manjunath, B. (2001). Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 800–810.

Lienhart, R. (2001). Reliable dissolve detection. *Storage and Retrieval for Media Databases*, **4315**, 219–230.

Miene, A., Hermes, T., Ioannidis, G. & Herzog, O. (2003). Automatic shot boundary detection using adaptive thresholds. *Proceedings of the TRECVID 2003 Workshop', Gaithersburg, Maryland, USA*.

Nagasaka, A. & Tanaka, Y. (1991). Automatic video indexing and full-video search for object appearances. *Proceedings of the IFIP TC2/WG 2.6 Second Working Conference on Visual Database Systems II*, 113–127.

Ngo, C., Pan, Z., Wei, X., Wu, X., Tan, H. & Zhao, W. (2005). Motion driven approaches to shot boundary detection, low-level feature extraction and BBC rush characterization at TRECVID 2005. *TRECVID proceedings*.

NIST (2007). National institute of standards and technology. *available at http://www.nist.gov (last seen on 20 August 2007)*.

Pickering, M. & Ruger, S. (2001). Multi-timescale video shot-change detection. *NIST Special Publication*, 500–250.

Pye, D., Hollinghurst, N., Mills, T. & Wood, K. (1998). Audio-visual segmentation for content-based retrieval. *5th International Conference on Spoken Language Processing, Sydney, Australia, Dec*.

Qi, Y., Hauptmann, A. & Liu, T. (2003). Supervised classification for video shot segmentation. *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, **2**.

Quenot, G., Moraru, D. & Besacier, L. (2003). Clips at TRECVID: Shot boundary detection and feature detection. *TRECVID 2003 Workshop Notebook Papers, Gaithersburg, MD, USA*, 18–21.

Ren, W. & Singh, S. (2004). Automatic video shot boundary detection using machine learning. In *IDEAL*, 285–292.

Smeaton, A.F., Over, P. & Kraaij, W. (2006). Evaluation campaigns and TRECVID. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, 321–330.

TRECVID (2007). Guidelines for the TRECVID 2007 evaluation. *available at http://www-nlpir.nist.gov/projects/tv2007/tv2007.html, (last seen on 20 August 2007)*.