# Multi-timescale video shot-change detection

Marcus J Pickering and Stefan M Rüger
Department of Computing
Imperial College of Science, Technology and Medicine
180 Queen's Gate, London SW7 2BZ, England
m.pickering@doc.ic.ac.uk

**Abstract**

We describe our shot-boundary detection experiments for the TREC-10 video track, using a multi-timescale shot-change detection algorithm.

## 1 Introduction

The shot change detection system is based on that proposed by Pye et al [1] as part of a scene-change detection method which also took into account audio cues.

The algorithm aims to detect and distinguish different types of breaks in video by looking at differences between frames across a range of timescales. Looking at a wider range of frames than just those that are consecutive enables the detection of gradual changes such as fades and dissolves, while rejecting transients such as those caused by camera flashes.

Influenced by Zhang's twin comparison method [2], we added functionality to detect the start and end times of gradual changes to fulfil the requirements for the TREC submission.

In the remainder of this paper, we describe the shot-change detection system in detail, and present the results of the TREC evaluation.

## 2 System Description

The video segmentation algorithm is broadly based on the colour histogram method, which is extended for detection of gradual transitions that take place over a number of frames, and for rejection of transients, such as the effect of a flash-bulb.

Each frame is divided into 9 blocks, and for each block a histogram is determined for each of the RGB components. The Manhattan distance between corresponding component histograms for each corresponding block in two images is calculated, and the largest of the three is taken as the distance for that block. The distance between two frames is then taken as the median of the 9 block distances. This helps eliminate response to local motion.

A difference measure is defined as follows:

$$d_n(t) = \frac{1}{n} \sum_{i=0}^{n-1} D(t+i, t-n+i),$$

where $D(i, j)$ represents the median block distance between frames $i$ and $j$.

| Video | Known Trans | Mean Recall | Recall | Mean Prec | Prec |
|---|---|---|---|---|---|
| ahf1.mpg | 63 | 0.961 | 0.936 | 0.919 | 0.921 |
| anni009.mpg | 38 | 0.870 | 0.789 | 0.735 | 0.697 |
| bor03.mpg | 230 | 0.856 | 0.982 | 0.920 | 0.953 |
| bor08.mpg | 379 | 0.920 | 0.873 | 0.894 | 0.945 |
| bor17.mpg | 127 | 0.905 | 0.952 | 0.843 | 0.975 |
| nad28.mpg | 181 | 0.917 | 0.939 | 0.853 | 0.762 |
| nad31.mpg | 187 | 0.892 | 0.866 | 0.834 | 0.900 |
| nad33.mpg | 189 | 0.951 | 0.952 | 0.882 | 0.918 |
| nad53.mpg | 83 | 0.962 | 0.951 | 0.876 | 0.797 |
| senses111.mpg | 292 | 0.902 | 0.989 | 0.909 | 1.000 |
| ydh1.mpg | 69 | 0.961 | 0.971 | 0.839 | 0.881 |
| Weighted means | | 0.912 | 0.932 | 0.879 | 0.916 |

Table 1: Results of shot-boundary detection task for cuts in all files with greater than 100 overall transitions. The recall and precision values for our system are shown next to the respective means across all systems. Recall is the proportion of the known shot boundaries retrieved by the system and precision is the proportion of boundaries retrieved by the system which were judged to be correct. The bottom line shows the column mean for each of the statistics, with each file's contribution weighted by the number of transitions in that file.

A peak is defined as a value of $d_n$ which is greater than a pre-defined threshold and is greater than the 16 preceding and 16 following values of $d_n$. A shot break is declared if there are near-coincident peaks of $d_{16}$ and $d_8$. An additional coincident peak of $d_2$ suggests a cut, otherwise the break is classified as a gradual transition.

The algorithm thus far detects the presence of cuts or gradual changes, but gives no indication of the start and finish points of the gradual changes. We therefore employ a method similar to that described by Zhang [2] in which a lower threshold is used to test for the start and end of a gradual transition. At each frame, the $d_4$ difference is compared to the threshold. If it is greater than the threshold it is marked as a potential start of a transition. If, on examination of successive frames, the $d_4$ difference falls below the threshold again before a shot change is detected, this potential start is scrapped and the search continues. Following the detection of a shot change, the end point of the transition is declared as the point at which the $d_4$ change first falls below the threshold again, following the shot change. The $d_4$ timescale is used because it is thought to be fine enough to accurately pinpoint the moment at which the change begins, but also introduces a tolerance to any momentary drop in the difference which may occur in the process of the change.

## 3 Results

The results show that our system performed slightly better than average overall. The breakdown of detected breaks into cuts (Table 1) and gradual transitions (Table 2) shows that, as for other systems, performance is considerably poorer for gradual transitions than it is for simple cuts. However, one of our best performances, in both precision and recall relative to the average, was

| Video | Known Trans | Mean Recall | Recall | Mean Prec | Prec |
|---|---|---|---|---|---|
| ahf1.mpg | 44 | 0.683 | 0.681 | 0.700 | 0.625 |
| anni009.mpg | 65 | 0.501 | 0.523 | 0.669 | 0.809 |
| bor03.mpg | 11 | 0.660 | 0.545 | 0.283 | 0.166 |
| bor08.mpg | 151 | 0.633 | 0.741 | 0.794 | 0.741 |
| bor10.mpg | 150 | 0.687 | 0.866 | 0.743 | 0.866 |
| bor12.mpg | 136 | 0.556 | 0.625 | 0.705 | 0.825 |
| bor17.mpg | 119 | 0.511 | 0.697 | 0.678 | 0.584 |
| nad28.mpg | 116 | 0.603 | 0.543 | 0.555 | 0.588 |
| nad31.mpg | 55 | 0.478 | 0.418 | 0.436 | 0.353 |
| nad33.mpg | 26 | 0.535 | 0.500 | 0.389 | 0.206 |
| nad53.mpg | 76 | 0.596 | 0.631 | 0.575 | 0.761 |
| senses111.mpg | 16 | 0.336 | 0.187 | 0.298 | 0.068 |
| ydh1.mpg | 52 | 0.492 | 0.423 | 0.620 | 0.536 |
| Weighted means | | 0.581 | 0.641 | 0.653 | 0.674 |

Table 2: Results of shot-boundary detection task for gradual transitions in all files with greater than 100 overall transitions.

for "bor12.mpg" for which all breaks were gradual changes.

On closer examination of the test videos, it becomes clear that the relatively poor performance of the system on gradual transitions is due in large part to the difficulty in distinguishing object or camera motion from gradual transitions. One of the lowest results for precision was obtained for the video "senses111.mpg", and examination of the video shows that almost all of the false positives were caused by the camera following the subject as he moved across a highly contrasting background, with the rest being caused by object motion. A look at the missed transitions shows that they were almost all situations in which one technical drawing was dissolved into another containing identical colours. To detect this would have required an extremely low threshold which would have brought with it a whole raft of new false positives.

One of the poorest recall rates for the results shown was for "anni009.mpg". Here the problem seems to be caused by a confusion between cuts and gradual changes - virtually all of the falsely declared cuts corresponded to missed gradual transitions, which suggests that the thresholds were inappropriately set for this recording.

Another cause of reduced performance was that our detection of transition start and stop times did not match with those of the reference. In "bor12.mpg", for example, a large number of the gradual transitions appearing in the list of inserted breaks correspond to longer transitions appearing in the list of deleted transitions. This may point to below-optimum threshold setting, but TREC's determination of the reference itself was a subjective process, therefore the disparity is not necessarily due to an inherent flaw in our system.

# 4 Conclusions

While our system's performance was, on balance, slightly better than average, there were some obvious problems. It is clear that some discriminant is needed in addition to colour. If an object could be tracked across a supposed shot boundary, the boundary could be discounted, for example. Other discriminants such as texture and shape may also be helpful.

There was evidence that the empirically determined thresholds were not optimal in all cases, suggesting that some form of automatic threshold adjustment may be required. However, this would probably require two passes over the data.

**Acknowledgements:** This work was partially supported by EPSRC, UK, and AT&T Research Labs, Cambridge, UK.

## References

[1] Pye D, Hollinghurst NJ, Mills TJ, Wood KR. Audio-Visual Segmentation for Content-Based Retrieval. 5th International Conference on Spoken Language Processing, Sydney, Australia, Dec 1998.

[2] Zhang HJ, Kankanhalli A, Smoliar SW. Automatic Partitioning of Full Motion Video. Multimedia Systems vol 1, 10-28, Jan 1993.