University of Sheffield at TRECVID 2008: Rushes Summarisation and Video Copy Detection

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

Abstract

This paper presents our experiments on TRECVID workshop 2008. This year we participated in two challenging tasks, rushes video summarisation and video copy detection tasks. We studied a spatio-temporal video model to represent inter-frame correlation between two streams of video. The rushes summarisation task is aiming at creation of summaries with length not more than 2% of the original video. The approach involves the following three steps: irrelevant frame sequence removal, clapper shot detection, and shot alignment. The evaluation is made by human judges in four categories. It indicates that created summaries do not contain many duplications and junks. The summaries also have pleasant rhythm. The purpose of the video copy detection task is to detect pairs of copy (query) and reference video from the collection. We studied the problem with audio only queries that utilises the spatio-temporal video model.

1 Rushes Video Summarisation

Rushes (or pre-production video) is a raw material that will be used to produce video data such as drama, and television programmes. Contents are normally unstructured and contain natural sounds, low quality audio video streams, and highly repetitive contents. The repetitive sequences are produced from the retake of the same scene. They are not copies; there exists some differences between shots caused by changes of a camera angle, or video production errors — *e.g.*, actors performing the incorrect line of a story, or some scene with missing objects (Over *et al.*, 2007). The nature of rushes indicates that we require sophisticated technologies for managing and accessing the unstructured contents.

In the rushes video task this year, summary clips are created, aiming at easy access and efficient use of the material (TRECVID, 2008). Summaries should include major objects and events specified. The objectives are (1) to minimise the number of frames in a summary video and (2) to present information in a way that maximises the usability. We investigate a spatio-temporal video model in order to align multiple shots (Section 1.2). The approach involves the following steps: firstly frame sequences with little information, such as colour bars and monochromatic frames, are removed (Section 1.1). All individual shots are then separated by detecting clapper boards (Section 1.3). The shot alignment is performed to group all repetitive shots (Section 1.4). Finally the summary is created by applying a set of criteria on the aligned shots. We submitted a single run for the test data set, which was evaluated by human judges.

1.1 Irrelevant Frame Removal

Colour bars and monochromatic frames are mostly irrelevant, yet frequently observed components in rushes video. Colour bars are often found at the beginning of video, prompting the start of some scene. They are also observed in the transition between shots. Occasionally monochromatic frames (*e.g.*, black, grey or white frames) appears perhaps due to a faulty camera. With the aid of uniqueness in their visual appearance these frames may be simply identified by detecting pixel discontinuities between two consecutive frames.

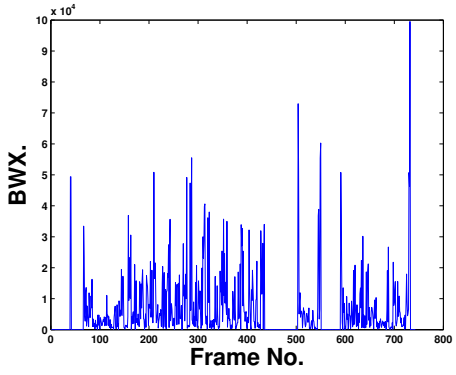


Figure 1: The calculation of BWX for a test video '*MRS044499*'. With 1 fps sampling rate, the graph contains 743 video frames. Intervals 1-68, 436-505, and 550-592 represents sequences of irrelevant frames, implying that there are three major chunks of video stories.

Shot boundaries are detected using the methodology we worked for TRECVID 2007 (Chantamunee & Gotoh, 2007). The black/white XOR (BWX) method operates the '*exclusive OR*' of every consecutive frames in grey scale. Figure 1 demonstrates the BWX calculation for the entire length of one test video. The approach can distinguish any irrelevant contents by identifying a gap in the figure.

1.2 Spatio-Temporal Video Model

After removing colour bars and monochromatic frames a single shot may contain several retakes of the same scene segmented by frames with a clapper board. A clapper board shows the scene number and the story name in order to maintain the story track that assists the production of the final video. Retakes duplicate contents of the same scene. Although identical objects exist, recordings may be made in different settings such as use of zoom or varieties of camera angles. Some production errors, such as incorrect line of actor's speech, may also lead to further retakes. Existence of many retakes results in a different summarisation procedure of rushes video from that of other kinds of video. Rushes video requires structuring of its repetitive contents in order to choose the most important scene.

The spatio-temporal model address the problem of aligning retakes by incorporating audio-visual information. The method utilises Gaussian mixture model (GMM) (Bishop, 1995) to model inter-frame correlation in order to synchronise multiple streams of repetitive shots. GMM can be applied for data clustering tasks. To model spatial and temporal information, their particular feature(s) can be extracted from every fixed-length cluster (*e.g.*, window of 1 second length) on rushes, from which inter-cluster correlation is calculated.

Technically, suppose that $x = \{x_1, x_2, \ldots, x_n\}$ represents a sequence of multi-dimensional feature vectors x_i. A spatial-temporal path C consists of a series of GMM clusters represented as

$$C = [c_1, c_2, \ldots, c_n] \tag{1}$$

where c_i is a set of GMM parameter derived from x_i. The expectation-maximisation (EM) algorithm is employed to estimate the GMM (Bishop, 1995). Each cluster represents the centroid of space at a particular time. Figure 2 visualises the spatial-temporal GMM path. The path is similar to a Markov chain model, but it does not maintain probabilistic state transition and instead preserves spatio-temporal property. The spatio-temporal information can be visualised as a matrix of probability density functions (pdf's) $f(x_i|C)$ for x_i with $i = 1, \ldots, n$. The matrix is shown in Figure 3.

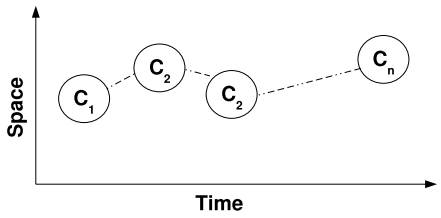
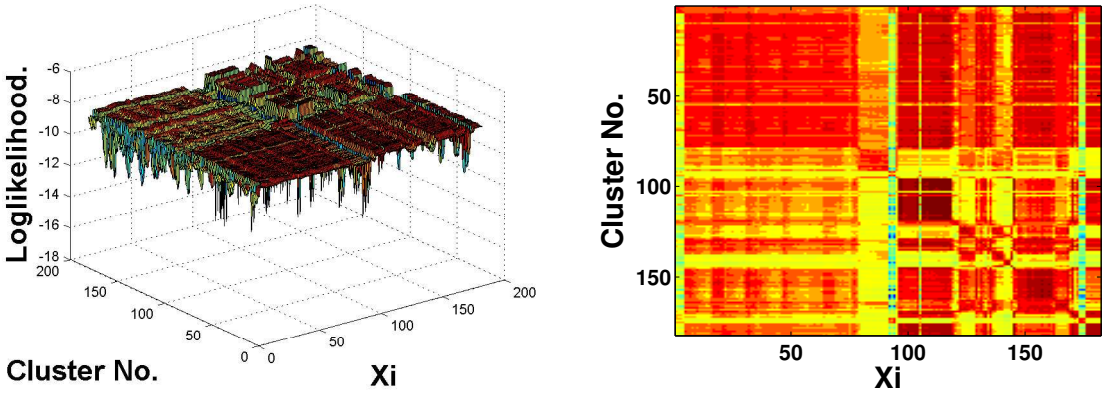


Figure 2: In the spatio-temporal path, clusters c_i are chained along timeline while their space is maintained.



(a) (b)

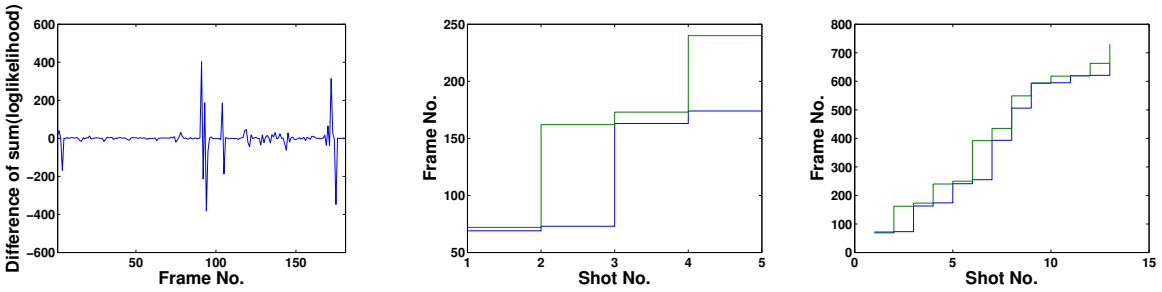
Figure 3: For a test video segment '*MRS044499*': a number of rectangular boxes can be observed along the diagonal line, representing space and time of each video shot.

1.3 Detection of Clapper Board

Frames with a clapper board frames have unique characteristics and appear with a large number of variations. A clapper board usually appears in a scene and stay for a moment. After performing shot tracking, the board is moved out from the scene (Pan *et al.*, 2007). The spatio-temporal video model can be applied to detect these frames. RGB colour histogram is employed as 256-dimensional visual feature vectors. Figure 4 demonstrates the approach, where shots are represented as rectangular boxes with frame numbers. Shots having short duration are considered not useful and removed. In the experiment, the threshold is set to 20 seconds, corresponding to 20 frames when using the frame rate of 1fps. They are mostly, if not all, frames of clapper boards.

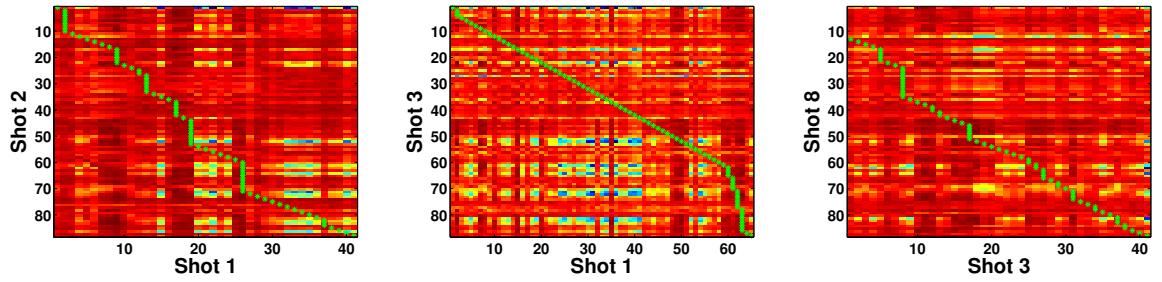
1.4 Shot Alignment for Redundancy Detection

We observe that a rushes video usually consists of one or two stories whose duration is about 5–10 minutes each. However retakes generate the duplication of the same scene, increasing the length of rushes. It therefore requires a method that selectively discards contents from the unstructured repetition. The recording of same scene is repeated due to several reasons described earlier. Although visual information in some retake scenes may differ, actors usually make a similar line of conversation. Our main idea for finding repetitive shots is to identify such similarity, if not the same, in conversation. The spatio-temporal video model can be applied to evaluate the auditory scene. The shots are then aligned in order to match the most likely line of conversation. To this end audio features are employed (as opposed to RGB colour histogram in Section 1.3). The alignment path of the length l is then traversed 'along the diagonal', which may not be 'on the diagonal', of the matrix. In practice, we operate the dynamic time



(a) (b) (c)

Figure 4: For a test video segment '*MRS044499*': panel (a) shows the difference between two consecutive frames after summing the log likelihood for every x_i in Figure 3(b). Each clear spike implies a shot boundary. Panels (b) and (c) show shot boundaries for a part and the entire length of video.



(a) (b) (c)

Figure 5: For a test video segment '*MRS044499*': alignment of (a) shots 1 and 2, (b) shots 1 and 3, and (c) shots 3 and 8. Note that shots have different lengths. Shots 1 and 3 are the best aligned pair, probably indicating that they contain the most complete detail of the scene.

wrapping (DTW) (Vlachos *et al.*, 2006) on the matrix to find the best alignment path between two shots. Let s=$\{s_1, s_2, \ldots, s_n\}$ and t=$\{t_1, t_2, \ldots, t_n\}$ be two comparative shots. Then,

$$\{\hat{i}_k, \hat{j}_k\}_{k=1,\ldots,l} = argmax_{i_k,j_k} \sum_{k=1}^{l} \log f(t_{j_k}|\theta_{s_{i_k}}) \tag{2}$$

gives the maximum (log)likelihood estimate for the shot alignment. Figure 5 illustrates three sample alignments. Aligned retakes are then grouped together, of which all continuous aligned contents are selected and ranked by the score of alignment.

1.5 Experiment and Result

NIST has provided a collection of BBC rushes video for documentary and TV drama. The collection consists of 97 video files in MPEG-1 format with the total length of 53 hours; 57 videos (35 hours) were used for development and the rest for testing. All of last year's rushes were used as this year's development collection (TRECVID, 2008).

In the experiment the video frame rate was set to 1 fps and the audio sampling rate is 44 kHz. The spatio-temporal video model was applied: RGB colour histogram was used for detection of clapper shots, and audio feature was used for shot alignment. For the latter, 13 MFCCs (mel frequency cepstral coefficients) were calculated from the audio track of the video. The Hamming window of 25ms was employed, overlapping 15ms with the adjacent windows. In the end, 1s of audio resulted in 13 dimensional feature

	shot alignment method	best	average	worst
pleasant tempo/rhythm (TE)	3.36	3.38	2.73	1.44
lots of duplicate (RE)	3.25	3.64	3.15	2.52
lots of junk (JU)	3.61	3.99	3.27	2.02
fraction of inclusions (IN)	0.17	0.83	0.44	0.07

Table 1: 40 video summaries were evaluated by three assessors in the following categories: summary has a pleasant tempo/rhythm (TE), summary contains lots of duplicate video (RE), summary contains lots of junk (JU), and fraction of inclusions found in the summary (IN). Score were given between 5 (strongly agree) and 1 (strongly disagree) for TE, between 5 (strongly disagree) and 1 (strongly agree) for RE, JU, and 1–0 (fraction) for IN. As a consequence, the higher score the better for all categories. Our approach was compared against the best, the worst, and the average of 42 international groups participating in the rushes video evaluation task this year.

vector of the length 100. The number of Gaussians in the mixture was set to five. Finally, all video fragments were simply concatenated to create a summary. In this task, the duration of summary clip was limited up to 2% of the original video.

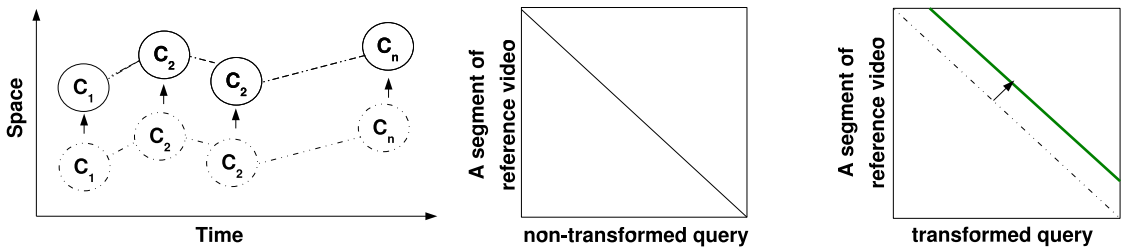
The evaluation was made by human judges. Three judges assessed the usability and quality of summaries. There were four questions; a summary had a pleasant tempo/rhythm (TE), contained lots of duplicate video (RE), contained lots of junk (JU), and fraction of inclusions found in the summary (IN). Questions could be answered in the range between 'strongly agree' and 'strongly disagree'. Table 1 indicates that our approach created a summary of pleasant rhythm without much duplication. It failed to capture many keyframes, containing approximately one out of five from the groundtruth. The groundtruth included the major objects from every scene even if it had a similar story. Therefore the weakness in inclusion was expected because the approach selected a single stream per scene from a video while the groundtruth was spread over the video.

2 Video Copy Detection

With the emerge of technologies for video capturing, storing, networking and digital media, the volume of video has been rapidly increasing and video authoring/editing becomes inexpensive. The digital media has become pervasively used for entertainment, education and business. Detection of copies of digital media (audio, image and text) has become important for managing the rapid increase of multimedia volume and protecting human innovation.

There are a wide range of works in recent years dealing with video copy detection. They are classified as two major approaches: watermarking and content-based video copy detection. For watermarking, authorised information (called watermark) is embedded into media prior to distribution. Later watermark can be extracted to check the authorisation (Judge & Ammar, 2002; Hzu & Wu, 1997). Content-based video copy detection does not insert any additional information. It employs traditional content-based information retrieval techniques. The unique information from video itself is used to detect the copies (Hampapur *et al.*, 2002; Zobel & Hoad, 2006). The latter technique has received wider attention due to many disadvantages associated with watermarking. It requires to ensure that distributed video has been watermarked which is not altered during capturing and transmission. This mechanism is costly for industries such as movie or television broadcast companies. On the other hand, content-based video copy detection is still a challenging task because of the great variation of techniques applied.

This year NIST launched the pilot task of content-based video copy detection (TRECVID, 2008). It aims at detection of video copies with various transformations. The transformation includes modification, addition, deletion of the original video by the aspect of colour, contrast, encoding, camcording, *etc.* The task is carried out with three different problems; image-based (required task), audio-based (optional),



(a) (b) (c)

Figure 6: A copy (query) is compared with a fragment of reference video. (a) the position of transformed segment is shifted with the constant amount. Panel (b) shows the diagonal alignment path for non-transformed query (which is not a copy). In panel (c), there is an alignment path with a constant distance from the diagonal line. This query is determined as a copy of the reference video.

and audio/visual-based copy detection (optional). We only worked on audio-based copy detection, therefore it was not evaluation by NIST. The following section discusses our approach and the experimental settings but without results.

2.1 Approach

For the TRECVID collection, we observe that transformations are made broadly in two different manners, either alteration of entire reference video or conversion on some parts of the video. In either of ways, the shifted values are changed with a constant value. The spatio-temporal model is employed for copy detection using audio features. Our assumption is that all changed values of query should make an alignment path shifting with the constant distance from the diagonal line. Figure 6 illustrates the idea when comparing a query to a segment of reference video. The query and the reference segment have the same length. For each combination of query and reference segment, the number of aligned frames are counted and used for ranking the best match.

2.2 Experimental Setup

The video dataset was a collection of 438 MPEG-1 video files. The training collection consisted of 100 hours of TRECVID 2008 Sound and Vision data. The combination of TRECVID 2007 search and HLF (High Level Feature) test data and another 100 hours of TRECVID 2008 Sound and Vision data was released as the test dataset. The video-only and audio-only transformation queries were developed by IMEDIA. Transformations were derived from a small segment of video either from the test collection or from unrelated videos. The 2000 video-only queries were derived with a range of transformations: camcording, picture-in-picture, insertion of pattern, strong re-recording, change of gamma, and combination of transformations. For audio-only queries, 201 queries were produced with alterations such as bandwidth limitation, other coding-related distortion (*e.g.*, sub-band quantisation noise), and variable mixing with unrelated audio content. Video/audio query was not provided, but the participants could combine video-only and audio-only queries using a general video modification tool such as *ffmpeg* (TRECVID, 2008). Our experimental setup was the same as alignment of rushes video, whereby 13 MFCCs coefficients were used for audio features.

3 Conclusion

We participated in TRECVID 2008 with the rushes video summarisation and video copy detection tasks. For rushes summarisation, it was found that rushes contained many repetitive contents which strongly

affected the quality of the summary. Our approach was based on the alignment of frame sequences. The assumption was that the best aligned frames should contain a good story line from which rushes summary could be extracted. We also saw the improved performance by removing colour bars and black/grey frames. According to assessment by human judges, our method achieves high scores regarding to pleasant summary creation and redundancy removal, but the result for inclusion of groundtruth was low.

Video copy detection task was launched this year. We carry out the experiment on the audio-only copy detection task. We applied the same method as redundant detection in rushes video summarisation. Our experiment was not evaluated, but we reported our approach and experiment settings in this paper.

References

Bishop. (1995) Neural networks for pattern recognition. Oxford University Press.

Chantamunee and Gotoh. (2007) University of Sheffield in TRECVID 2007: Shot boundary detection and rushes video summarisation. Notebook paper, TRECVID workshop.

Hampapur, Hyun, and Bolle. (2002) Comparison of sequence matching techniques for video copy detection. SPIE Conference on Storage and Retrieval for Media Databases.

Hzu and Wu. (1997) Digital watermarking for video. Digital Signal Processing Proceedings.

Judge and Ammar. (2002) WHIM: watermarking multicast video with a hierarchy of intermediaries. The Journal of Computer Networks, 39, 6.

Over, Smeaton, and Kelly. (2007) The TRECVID 2007 BBC rushes summarization evaluation pilot. ACM Workshop on TRECVID Video Summarization, Augsburg, Germany.

Pan, Chuang, and Hsu. (2007) NTU TRECVID-2007 fast rushes summarization system. ACM Workshop on TRECVID Video Summarization, Augsburg, Germany.

Guidelines for the TRECVID 2008 evaluation.
Available at *http://www-nlpir.nist.gov/projects/tv2008/tv2008.html*.

Vlachos, Hadjieleftheriou, Gunopulos, and Keogh. (2006) Indexing multidimensional time-series. The VLDB Journal 15, 1.

Zobel and Hoad. (2006) Detection of video sequences using compact signatures. ACM Transactions on Information Systems, 24.