A GENERIC NEWS STORY SEGMENTATION SYSTEM AND ITS EVALUATION

Neil O'Hare, Alan F. Smeaton, Csaba Czirjek, Noel O'Connor, Noel Murphy

Centre for Digital Video Processing,
Dublin City University, Glasnevin, Dublin 9, Ireland
nohare@computing.dcu.ie

ABSTRACT

This paper presents an approach to automatically segmenting broadcast TV news programmes into individual news stories. We first segment the programme into individual shots, and then a number of analysis tools are run on the programme to extract features to represent each shot. The results of these feature extraction tools are then combined using a Support Vector Machine trained to detect anchorperson shots. A news broadcast can then be segmented into individual stories based on the location of the anchorperson shots within the programme. In this paper we use one generic system to segment programmes from two different broadcasters, illustrating the robustness of our feature extraction process to the production styles of different broadcasters.

1. INTRODUCTION

Over the last decade the problems involved with the capture, compression, streaming, transmission and playback of digital video information have largely been solved. This has meant that it is now possible to deploy digital video on a large scale. Once video content becomes available in very large quantities, however, it becomes important to be able to access this content as efficiently as possible.

This situation has led to much research in the areas of multimedia analysis and retrieval. The Centre for Digital Video Processing (CDVP) at Dublin City University has been pursuing an ongoing research programme to develop technologies for the efficient management of digital video content. The work of the Centre is demonstrated via the web-based Físchlár systems, which include the dedicated TV news sub-system Físchlár-News [1].

There has been much work in video processing in the area of temporal segmentation since an important task in managing digital assets is to break them down into manageable units. Often this segmentation is at the level of the shot, the task being to detect the boundaries between different camera set-ups. A summary of the video can then be presented based on this segmentation [2]. This shot-level segmentation can be understood as the detection of syntactic boundaries within digital video files, but in many cases a semantic segmentation, where higher level scene or story level changes are detected, is more useful. Detecting semantic boundaries is a difficult task, but in the case of news broadcasts we can use a priori knowledge of the structure of the broadcast to guide us in our segmentation (eg [3]).

In recognition of the fact that news broadcasts are an important genre, and their segmentation an important task, a Story Segmentation task was included as part of TRECVID2003 [4].

In this paper, we outline our own work in the Story Segmentation task for TRECVID2003. In the next section we outline the motivation for our approach and give an overview of our system. Section 3 describes the analysis tools we run on news broadcasts to extract features to be used for anchorperson detection. Section 4 describes how these features are combined to detect the anchorperson shots in a news broadcast using a Support Vector Machine. In Section 5 some results are presented and discussed, and in Section 6 some conclusions and directions for our future research are outlined.

2. SYSTEM OVERVIEW

In our work on News Story Segmentation, we exploit the structure of news broadcasts in order to detect story boundaries. Each story within a broadcast usually begins with a leading shot of the anchorperson introducing the new story. This anchor shot is usually followed by a report with more details on the story. Anchor shots are typically filmed in a studio location, and within a single broadcast they are captured via the same camera set-up, and as such exhibit strong visual similarity. This structure is illustrated in Figure 1.

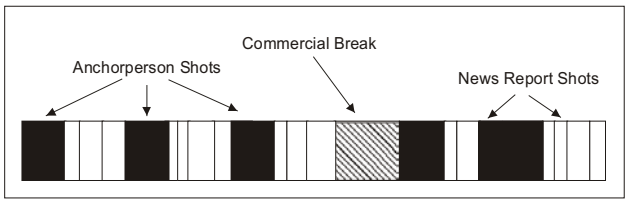

Figure 1. Structure of a News Broadcast

Our approach, outlined in Figure 2, is to detect the anchorperson shots by combining the results of a number of automatic video analysis tools: we make the assumption that all anchorperson shots signal a new story so we log stories at the beginning of anchorperson shots. We first segment the broadcast into shots, and subsequent analysis takes place at the level of the shot. We run a number of analysis tools to extract high-level features describing each shot in the video sequence, and then we combine these features using a trained Support Vector Machine (SVM). The resulting trained SVM classifier can be used to detect the anchorperson shots within a broadcast.

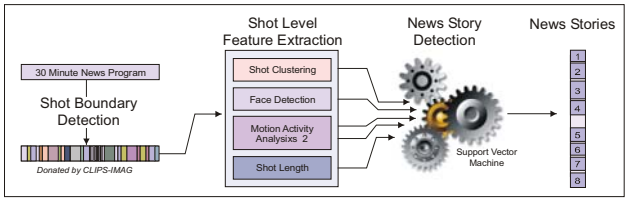

Figure 2. Overview of News Story Segmentation system.

3. VISUAL ANALYSIS FOR PRE-PROCESSING AND FEATURE EXTRACTION

The first stage of our news story segmentation approach is to segment each news broadcast into individual shots. For the TRECVID2003 story segmentation task, we used the common shot boundaries provided by the CLIPS-IMAG group [4] and used by all TRECVID participants. We then analyze each shot in the broadcast to extract a number of features representing the shot, as described below.

3.1. Shot Clustering

We have previously developed an approach for detecting anchorperson shots in a news programme based on shot clustering [5].

This method works by grouping together shots that are visually similar to each other, based on colour histograms. The Cosine similarity measure is used to calculate a distance between the keyframes representing each shot, and shots are clustered based on this distance.

Since the anchorperson shots in a TV news broadcast are very similar to each other, they are placed in the same cluster by this process. We can then apply some heuristics to distinguish the anchorperson groups from the non-anchorperson groups:

- The temporal range of the shots must be higher than a pre-determined threshold. Anchorperson shots tend to be dispersed throughout a broadcast, so shots that are very similar visually but occur only very close together in the broadcast should be rejected.
- The group similarity mean should be higher than a very high threshold. This is because anchorperson shots are extremely similar to each other and form 'tight' clusters.
- The mean anchorperson shot length should be longer than a minimum threshold. This is because anchorperson shots are generally quite long in comparison with other video content, rarely lasting less than five seconds.

This process gives a binary output: either a shot belongs to an anchorperson group or it does not belong to an anchorperson group. We converted this into a confidence value in the interval between 0 and 1 as follows. After the candidate anchorperson clusters have been identified, each keyframe is compared to each of the anchorperson groups using the Cosine similarity measure, giving a visual similarity between each keyframe and each anchorperson group. The maximum of these values can then be seen as a confidence that a given keyframe belongs to an anchorperson cluster, and this is output from the clustering algorithm for each shot.

3.2. Face Detection

The CDVP has also developed an approach for automatically detecting faces in digital video sequences [6]. An anchorperson shot in a news programme will always contain faces, and this information is valuable for an automatic news story segmentation system. The face detection process is summarised in Figure 3.

The first step in the face detection process involves colour analysis of the images. Since the colour of human skin falls within a relatively narrow band of the colour spectrum, it is possible to detect skin-like pixels. Morphological filtering is then used to obtain smoothed homogeneous areas of connected pixels. Shape and size heuristics remove some of these candidate regions. Any candidates remaining are rescaled to 64 x 64 pixels and passed to a principal component analysis (PCA) module.

The PCA process examines candidate regions and classifies them as faces or non-faces. This approach uses a number of training images to form a set of basis vectors, known as Eigenfaces, spanning an optimal subspace of the input space such that the mean square error between the projection of the training images onto this subspace and the original training images is minimised. Candidate face regions can now be analysed using these Eigenfaces via two distinct distance measures. Distance from Face Space

(DFFS) measures the distance between a test image and its reconstruction using only the first *N* Eigenfaces, and Distance in Face Space (DIFS) measures the distance between a test image and the average face. Both of these distance measures can be combined to give a confidence value indicating whether an image is a face or not.

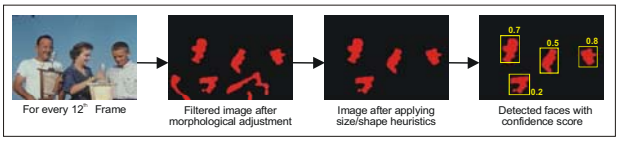

Figure 3. The face detection process.

This analysis is carried out on every 12th frame in each shot, and the confidence value for the presence of a human face in a shot is the average confidence value for all frames processed within that shot.

3.3. Motion Activity Analysis

Since anchorperson shots tend to have low visual activity with the only motion usually corresponding to the anchorperson's head/body/mouth moving, we measure visual activity across frames to aid our anchorperson detection. To measure visual activity in a shot, we analyse the motion vectors in the MPEG-1 bitstream. Our approach is similar to that of Sun et al [7], extending the approach to represent the motion over a shot rather than only in individual P-frames. We count the number of motion vectors in the frame whose length is less than a threshold. Only the motion vectors for the P-macroblocks are considered, since I-blocks have zero length motion vectors but do not represent zero motion. An activity measure is calculated as the ratio between the number of these short length P-blocks and the total number of macroblocks in a frame.

We use two separate techniques to extend this approach to extract a measure to represent the activity over a shot. For the first approach, the P-frame with the least amount of activity is used to represent the shot. This increases the likelihood of a shot being represented as having low activity: in the context of anchorperson detection this means that we are less likely to rule out candidate anchor shots because noise in the data makes them seem more active than they really are.

A second approach for representing the motion within a shot is to sum the motion vectors in all P-frames in a shot to represent the overall movement. A cumulative motion vector for each macroblock location is calculated as the sum of the motion vectors for all P-frames in the shot at that location, as illustrated in Figure 4. In an anchorperson shot the only motion tends to be the mouth or possibly the body movement of the anchorperson, which cancels itself as the anchorperson leans first in one direction and then another. This motion should be well represented by cumulative motion vectors, with motion vectors pointing in opposite directions cancelling each other out over the duration of anchor shots. We can calculate an activity measure for the entire shot using these cumulative motion vectors in the same way as for a normal P-frame.

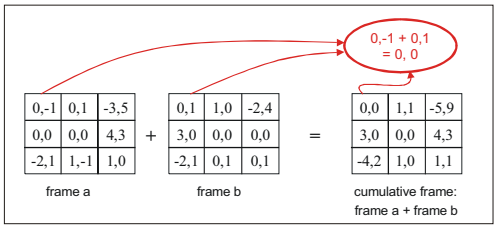

Figure 4. Calculating Cumulative Motion Vectors.

4. COMBINATION OF EXTRACTED FEATURES USING A SUPPORT VECTOR MACHINE

Support Vector Machines (SVMs) are a recently developed machine learning algorithm [8]. They have proven very popular in pattern recognition and have shown good performance in a variety of application areas. The approach works by finding the separating hyperplane between input classes that has the maximum distance from each class. This reduces the risk of error when classifying previously unseen test examples. It has the feature that the data is only represented as a dot product between input points. This allows for the dot product to be replaced by a kernel function (the dot product in a feature space) to cater for datasets with a non-linear separation.

For our work in news story segmentation, we used an SVM to combine the outputs of the feature analysis described in Section 3. From the TRECVID2003 development set [4], a subset of 20 programmes was selected as the training set, representing 10 hours of video footage. Each shot within these programmes is represented by a vector of the extracted features. A shot was labeled as a positive example if the shot contains a story boundary, and negative otherwise. In addition to the features described, the length of each shot, measured in frames, is also input to the SVM, since anchorperson shots introducing a new story tend to be longer than others.

The resulting SVM classifier, trained using the publicly available SVM-light system [9], can be used to detect the anchorperson shots in a broadcast. We then log story boundaries as the beginning of these anchorperson shots.

5. EVALUATION

We evaluated our system against the TRECVID2003 test collection, which had been manually marked up for story boundaries, using the standard evaluation measures of precision and recall. The TRECVID2003 collection

contains broadcast TV news from CNN and from ABC. In order to test the robustness of our feature extraction to variations in the production styles of different broadcasters, we trained a number of systems and tested each against the complete TRECVID2003 test set (both CNN and ABC):

1. A generic system was trained using both CNN and ABC material and evaluated on a collection consisting of both CNN and ABC.
2. A system was trained using only ABC data and was evaluated on a collection of both CNN and ABC.
3. A system was trained using only CNN material and was evaluated on a collection of both CNN and ABC.
4. We also tested the shot clustering algorithm run in isolation as a story bound detector. For this system any candidate anchor shots identified by the clustering algorithm are logged as story bounds, without including the other extracted features.

Initial results were poor (0.33 recall and 0.43 precision) but by adding a random element to the clustering process we were able to achieve an improvement in precision and recall. The results for each system are shown in Table 1.

All combined systems outperformed the clustering algorithm run in isolation, giving a good increase in precision for a relatively moderate loss in recall, clearly showing that combining features using a Support Vector Machine does give an improvement in performance. Surprisingly, each of the systems trained on a single broadcaster (systems 2 and 3) actually outperform the system trained on both broadcasters (system 1): this is something we intend to investigate further at a later stage.

System	Recall	Precision
1	0.4298	0.5349
2	0.4647	0.5193
3	0.4794	0.5110
4	0.519	0.4030

Table 1. Results for News Story Segmentation system against the TRECVID2003 test set.

6. CONCLUSIONS AND FUTURE WORK

The results presented in the previous section are slightly disappointing, particularly considering that other participants in the TRECVID2003 Story Segmentation task achieved 0.71 precision and 0.76 recall for runs involving audio visual analysis only (i.e. no text analysis). This disparity can be partly explained by the fact that our system uses no audio features, only video, unlike other systems it is being compared with, and our system used less visual features than others. Our main aim has been the development of a generic segmentation system that is robust to variations in production styles between different broadcasters. The fact that we can train our system on one broadcaster and test on another without any loss of performance is very encouraging in this respect. The news story segmentation system described in this paper has been running daily as part our Físchlár-News system [1] and has recently shown its versatility when the national TV broadcaster changed its TV news production style with new studio setting, etc.

Our system is clearly dependent on the results of the shot clustering algorithm and the other features essentially act as filters to its output. We plan to make this algorithm more robust by using more features than just colour histograms to represent shots in the clustering process. Also, we plan to introduce more diverse feature extraction tools, so that our system is not so dependent on one algorithm. This includes text-based segmentation using closed captions, and in the audio analysis domain we will incorporate a speaker segmentation tool and a silence detection tool.

Acknowldgements. This material is based upon work supported by the IST programme of the EU in the project IST-2000-32795 SCHEMA.

7. REFERENCES

[1] Smeaton, A.F., Murphy, N., O'Connor, N., Marlow, S., Lee, H., McDonald, K., Browne, P., Ye, J. "The Físchlár Digital Video System: A Digital Library of Broadcast TV Programmes", *JCDL 2001, ACM + IEEE Joint Conference on Digital Libraries,* Roanoke, VA, 24-28 June 2001.

[2] Zhang, H.J., Low, C., Smoliar, S., Wu J. "Video parsing, retrieval and browsing: an integrated and content-based solution", *Proceedings of ACM Multimedia*, New York, 1995.

[3] Zhang, H., Gong, Y., Smoliar, S., Tan, S. "Automatic Parsing of News Video". *Proceedings of the IEEE International Conference on Multimedia Computing.and Systems,* 1994.

[4] Guidelines for the TRECVID 2003 Evaluation. http://www-nlpir.nist.gov/projects/tv2003/ (last visited January 2004)

[5] O'Connor, N., Czirjek, C., Deasy, S., Marlow, S., Murphy, N., Smeaton, A.F. "News Story Segmentation in the Físchlár Video Indexing System", *Proceedings of the International Conference on Image Processing (ICIP 2001)*, Thessaloniki, Greece , 10-12 October 2001.

[6] Czirjek, C., O'Connor, N., Marlow, S., Murphy, N. "Face Detection and Clustering for Video Indexing Applications", *ACIVS 2003 - Advanced Concepts for Intelligent Vision Systems,* Ghent, Belgium , 2-5 September 2003.

[7] Sun, X., Manjunath, B.S., Divakaran, A. "Representation of motion activity in hierarchical levels for video indexing and filtering", Proceedings of IEEE International Conference on Image Processing (ICIP), Rochester, NY, USA, September 2002.

[8] Burges, C., "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery*, 2(2):121-167, 1998.

[9] Joachims, T. "Making large-Scale SVM Learning Practical", *Advances in Kernel Methods - Support Vector Learning*, Schölkopf , B., Burges, C., Smola, A., (eds.), MIT-Press, 1999.