VIDEO GENRE CLASSIFICATION USING DYNAMICS

M.J. Roach, J.S.D. Mason

Department of Electrical
& Electronic Engineering
University of Wales, Swansea
SA2 8PP, UK
eeroachm@swansea.ac.uk
http://galilee.swan.ac.uk/

M. Pawlewski

BTexaCT
Advanced Communication Technologies
Adastral Park
Martlesham Heath
Ipswich IP5 3RE

ABSTRACT

The problem addressed here is classification of videos at the highest level into pre-defined genre. The approach adopted is based on the dynamic content of short sequences (∼30 secs). This paper presents two methods of extracting motion from a video sequence: foreground object motion and background camera motion. These dynamics are extracted, processed and applied to classify 3 broad classes: sports, cartoons and news. Experimental results for this 3 class problem give error rates of 17%, 8% and 6% for camera motion, object motion and both combined respectively, on ∼30 second sequences.

1. INTRODUCTION

There is a substantial amount of multimedia data in the world, diverse in content and origin. In order to make efficient use of this data it should be labelled or indexed in some manner. Such labelling would make it easier for individuals to retrieve the type of material desired. In particular the area of multimedia data considered here is the leisure broadcast type i.e. T.V. programs, films etc. The task is to supply customers with material of their viewing preference. In this paper we look at the classification of broadcast material at the highest level of genre namely: sports, news and cartoons.

Although broadcast material can be labelled at the production stage there is still a need for automatic classification of videos. First, lots of videos currently exist that to date have not been labelled. Second, and perhaps most importantly, content-based classification approaches are the ultimate filter especially for broadcasting. Unlike labels and water marks which are susceptible to human error and fraud, content-based approaches are dependent solely on the actual material. The only limit of content-based approaches is the accuracy of the system itself. Furthermore the system can be considered as a final check, and complementary in providing useful additional meta-data to any static labelling system.

There are many approaches to content-based classification of videos ranging from low-level, limited environment, event detection through to high-level, broad environment genre classification. These approaches can be based on static features, dynamic features or a combination of the two.

If the environment is limited then the classification task can become more specific as for example in [1] were semantic attributes of captions are used for classification of news videos. A combination of static and dynamic features used in a limited environment is presented by Haering *et al* in [2] where event detection is applied to detecting hunts in wildlife videos. Another combination approach is applied to sports sequences by Yow *et al* in [3] who analyse soccer video for highlights, where the ball is tracked and the static up-rights of the goal posts are detected to indicate a shot on goal. These applications require constrained inputs for success; they rely on the video being pre-classified into news, wildlife and sport respectively. It is this high level of video classification to which our approach is applied.

Other approaches applied to this high level of video classification, where the input environment is relatively unlimited, can be generalised into two broad categories. There are those techniques that separate the video into previously defined categories and those that are example-based query tools.

In the later case, often termed query by example, the system is presented with an example or small number of examples of the type of video sequence required. The system then searches a database for videos with similar attributes. These approaches can vary in functionality and complexity of the similarity measure between the example and retrieved video sequences. An automatic approach by Chang *et al* supporting spatio-temporal queries is presented in [4]. A survey on content-based video retrieval is presented by Yongsheng *et al* in [5].

Within the case where the classes are pre-defined, there is a great deal of variation. First, there is variation in the classes or genre of video, in number and type. Some popular categories are sports, news, cartoons and commercials. An approach that includes all these categories plus music videos is presented by Dorai *et al* in [6]. A number of holistic global features including statics and dynamics are used for classification. Obviously, the most successful approaches will incorporate feature extraction from the audio signal as well as the video. A good example of this is presented by Fischer *et al* [7] using image based and audio based features applied to video genre classification. The approach here presents dynamic feature extraction applied to 3 predefined classes.

Approaches that are based on only dynamics include those ranging from accurate tracking of object motion to more crude region-based motion measures. Event classifiers usually have complicated low-level motion measures, as described by Courtney [8]. Moving objects are tracked and representations of their movements are then classified to identify events such as motion/rest, entrance/exit of objects etc. Similar motion descriptors, have been developed for object motion by Jeannin *et al* [9] along with camera motion descriptors and are presented for application for the MPEG 7 standard. These approaches are tend to be relatively computationally intensive.

Dynamic approaches that use less complex motion measures to classify video sequences are presented by Bouthemy *et al*. For example in [10, 11] local motion measures and global motion features are used to classify temporal textures such as fire and foliage; they also claim that these measures can be used to retrieve clips of similar global motion properties such as sports.

Here we investigate the performance of a purely dynamic based approach [10, 11] applied to video classification as the works of [9] and comparable to [6]. The remainder of the paper is structured as follows. In Section 2 the particular video dynamics and motion extraction are defined. In Section 3 the experimental conditions and results are presented leading to conclusions in Section 4.

2. VIDEO DYNAMICS

There are three different identifiable types of dynamics in video sequences. Two are contained within the scene, namely the background or camera motion and the foreground object motion. The third is the rate of shot or scene changes and is of a different origin, namely external manual editing influences. Although this third form of dynamic has been reported in [6, 7] to have discriminatory properties within the application to video classification, it has not been considered here although it is the subject of further work. The following sections describe the motion feature extraction for the two dynamics considered. The background or camera motion is covered in Section 2.1, and a more detailed account can be found in [12]. The foreground object motion is covered in Section 2.2. See also [13].

2.1. Background camera motion extraction

Here a simple three-parameter motion model is used to deal with camera motion X, Y and Z. Left and right, or X motion includes X translation and pan. Up and down, or Y motion includes Y translation and tilt, and Z motion includes Z translation and focal length modifications or zoom. This provides relative computational efficiency and is predicted to include most of the discriminatory information useful for high-level classification.

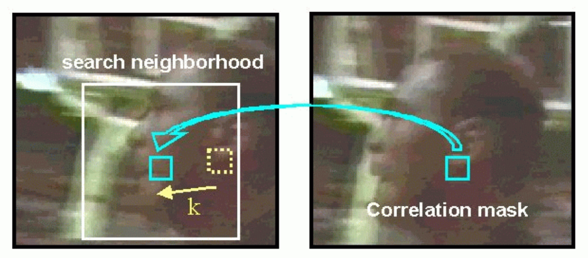

Fig. 1. Neighborhood search for similar blocks

In order to extract camera motion an optical flow based approach has been adopted. A test mask, labelled correlation mask in Figure 1, is extracted from the current image and the previous image is searched to find a similar mask. The translation of the block is usually small and therefore, to decrease computation, only a fraction of the previous image is searched, called the search neighborhood, also shown in Figure 1. When the most similar block in the search neigborhood is found, two thresholds are checked: one against the corelation coefficent to make sure the block is similar enough to be assumed the same block, two the number of similar enough blocks within the neigborhood is counted, if there are too many this indicates a uniform area and the motion vector k is unreliable. When both these criterior are met then the motion vector, k, can then be determined good. The similarity measure used to find the motion of the optical flow block is based on correlation and given by:

$$\mathfrak{C}_{l,m} = \frac{\sum_{i,j,c} P_c(i,j) Q_c(i+l, j+m)}{\sqrt{\sum_{i,j,c} P_c(i,j)}\sqrt{\sum_{i,j,c} Q_c(i+l, j+m)}} \quad (1)$$

where (l, m) locates the N×N search neighborhood and defines the center of the test masks, (i, j) describes the mask n×n, $c \in \{R, G, B\}$ is the red, green or blue, P_c is the colour value of the pixel in the current image, and Q_c is the colour value of the pixel in the previous frame.

This process is repeated over the image to obtain a full optical flow. Large homogeneous regions, typically half of the image, are considered to be the background. The camera motion is calculated to be the average motion vector $\bar{k}$ of the blocks contained within this background region. The camera $\bar{k}$ has X, Y and Z motion components. A second order signal of the X and Y components of $\bar{k}$ are calculated to create camera motion signals M_x and M_y.

2.2. Foreground object motion

To measure the foreground object motion first the camera motion is subtracted from the scene. Figure 2 shows the foreground motion extraction. The two original frames can be seen in Figures 2(a) and 2(b).

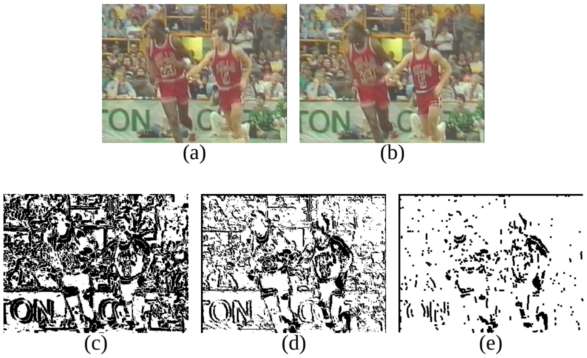

Fig. 2. Motion in a scene: (a) frame t-1, (b) frame t, (c) original, (d) camera compensated, (e) morphologicaly filtered

The motion is extracted by pixel-wise differencing of consecutive frames using the equation:

$$P(t) = P_{t(x,y)} - P_{t-1(x,y)} \begin{cases} 1 & > \text{threshold} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $P(t)$ is the Euclidean distance between colour pixels (R,G,B) in consecutive frames t and $t-1$, and x,y represent the pixel location. The motion between the frames is shown by black pixels. The result in Figure 2(c) is of all the motion within the scene. The camera compensation removes most of the background motion, Figure 2(d). To further enhance the object motion morphological opening is applied, the result of which is shown in Figure 2(e). Here the motion of the objects, basketball players, is most concentrated around the arms and legs. From a sequence of enhanced difference frames, such as 2(e), a second order object motion signal δ_t is calculated using the equation:

$$\delta_t = \frac{\sum_{x=1}^{x=w} \sum_{y=1}^{y=h} P(t)}{w \times h} dt \quad (3)$$

where w and h are width and height of the image respectively and $P(t)$ is given in Equation 2. Examples of these signals for each class can be seen in Figure 3.

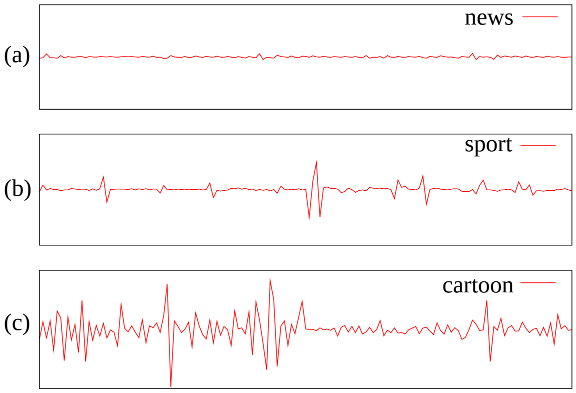

Fig. 3. Object motion signals for: (a) news, (b) sport and (c) cartoon

These signals represent the magnitude of the rate of change in the motion of the foreground objects. As may be expected news 3(a) has relativley little foreground motion and cartoon 3(c) has the most. This signal δ_t for each class along with the camera motion signals M_x and M_y are further processed for classification.

3. EXPERIMENTS

The modeling of the signals is as described in detail in [13]. The magnitude spectra of the second order motion signal δ and the camera motion signals M_x, M_y are processed using a DCT to provide low-pass filtering, orthogonality and a reduced feature dimension. The first n coefficients of each signal representation are concatenated into a single feature vector. The feature vectors are then used to train a Gaussian Mixture Model (GMM) based classifier [14].

The experiments are based on a subset of the database previously reported in [13, 12], comprising of a total of 18 sequences: 8 sport, 8 cartoon and 2 news each about 50 seconds long. A round-robin technique is used to maximise the use of the data: training on 14 minutes of video and testing on 1 minute then rotating the sequences. Assessment is performed on randomly chosen segments; the experiment is repeated for different length segments and classification performance assessed.

3.1. Experimental objectives

The objectives are to demonstrate the discriminatory properties of the two types of video dynamics:

- first, the camera motion X and Y.

- second the rate of change of the foreground object motion.
- finally the combination at feature level of the two types of video dynamics.

3.2. Results

Figure 4 shows classification error rate (%) against length of sequence (seconds). The 3 profiles show, camera motion, object motion and the two combined. The highest profile shows that the camera motion alone is the weakest discriminator. The best identification error for the camera motion of 17% is at ∼30 seconds. The profile for the foreground object motion shows much more discrimination. Here the error rate falls from 42% at 3 seconds to level off around 8% at ∼30 seconds. However, when added together at a feature level, the camera motion improves the results, giving the system its best performance of 6% error at ∼30 seconds. This performance is comparable with that of Dorai *et al* [6] who report approximately 90% accuracy using both static and dynamic features to classify sport, cartoon, news, commercials and music videos, obtaining the best results on 60 seconds of video on a 5 class problem.

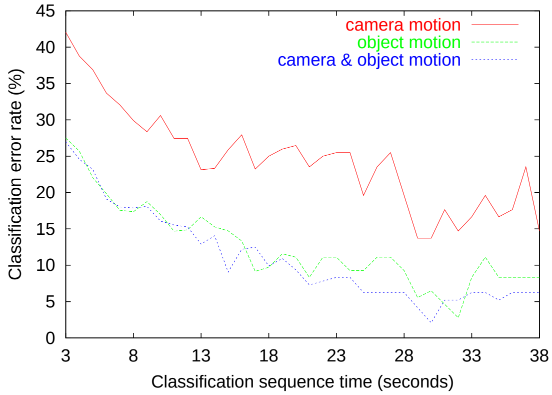

Fig. 4. Identification performance

4. CONCLUSIONS

The discriminatory properties of different types of dynamics within video sequences have been presented. The motion measures are content-dependent and are therefore the ultimate filter, in that they are based on the observed content. Thus they can be complementary to any form of static labeling e.g. meta-data. The results show that the dynamic feature extraction methods reported have good discriminatory properties and justify being part of an overall classification system possibly including static and audio features, such as in [6, 7]. Using just ∼30 second randomly chosen clips the system has a classification error rate of about 6% applied to the 3 video classes sport, cartoon and news.

5. ACKNOWLEDGEMENTS

M. Roach, wishes to express his gratitude to EPSRC and BT for their financial support. The authors also thank, BT and UWS research group members for their inputs.

6. REFERENCES

[1] I. Ide, R. Hamada, H. Tanaka, and S. Sakai, “News Video Classification based on Semantic Attributes of Captions,” in *Proc. 6th ACM International Conference*, 1998, pp. 60–61.

[2] N.C. Haering, R.J. Qian, and M.I. Sezan, “A Semantic Event Detection Approach and Its Application to Detecting Hunts in Wildlife Video,” *IEEE Trans. on Circuits and Systems for Video Tecnology*, 1999.

[3] D. Yow, B-L Yeo, M. Yeung, and B. Liu, “Analysis and presentation of soccer highlights from digital video,” in *Proc. Asian Conf. on Computer Vision*, 1995.

[4] S-F. Chang, W Chen, H. J. Meng, H. Sundaram, and D. Zhong, “A Fully Automated Content Based Video Search Engine Supporting Spatio-Temporal Queries,” *IEEE Trans. CSVT*, vol. 8, no. 5, pp. 602–615, 1998.

[5] Y. Yongsheng and L. Ming, “A Survey on Content based video retrival,” *http://www.cs.ust.hk/faculty/dimitris/COMP530/video_survey.pdf*.

[6] B-T. Truong, S. Venkatesh, and C. Dorai, “Automatic Genre Identification for Content-Based Video Categorization,” *Int. Conf. Pattern Recgnition*, vol. 4, pp. 230–233, 2000.

[7] S. Fischer, R. Lienhart, and W. Effelsberg, “Automatic Recognition of Film Genres,” in *The 3rd ACM Int. Multimedia Conference and Exhibition*, 1995.

[8] J. D. Courtney, “Automatic video indexing via object motion analysis,” *Pattern Recognition*, vol. 30, no. 4, pp. 607–625, 1997.

[9] S. Jeannin, R. Jasinschi, A. She, T. Naveen, B. Mory, and A.Tabatabai, “Motion descriptors for content-based video representation,” *Signal Processing Image Communication*, vol. 16, pp. 58–85, 2000.

[10] R. Fablet and P. Bouthemy, “Motion-Based Feature Extraction and Ascendant Hierarchical Classification for Video Indexing and Retrieval,” *3rd Int. Conf. on visual Information Systems, VISual’99, Amsterdam*, 1999.

[11] P. Bouthemy and R. Fablet, “Motion Characterization from Temporal Co-occurences of Local Motion-based Measures for Video Indexing,” *14th Int. Conf. on Pattern Recognition, ICPR’98, Brisbane*, 1998.

[12] M.J. Roach, P. Martin-Granel, and J.S.D. Mason, “Camera Motion Extraction using Correlation for Motion-Based Video Classification,” in *Submmited to Int. Workshop on Visual Form, Capri, Italy*, 2001.

[13] M.J. Roach and J.S.D. Mason, “Motion-Based Classification of Cartoons,” *Submmited to ISIMP, Hong-Kong*, 2001.

[14] D. A. Reynolds, “A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification.,” *Ph.D. thesis, Georgia Institute of Technology*, Sept. 1992.