AUTOMATED VIDEO SEGMENTATION

WEI REN, MONA SHARMA, SAMEER SINGH

PANN RESEARCH, DEPARTMENT OF COMPUTER SCIENCE,
UNIVERSITY OF EXETER, EXETER EX4 4PT, UNITED KINGDOM
{w.ren, m.sharma, s.singh}@exeter.ac.uk

ABSTRACT

Video segmentation is the process of dividing a sequence of frames into smaller meaningful units that represent information at the scene level. This process serves as a fundamental step towards any further analysis on video frames for content analysis. In the past, several statistical methods that compare frame differences have been published in literature and a range of similarity measures between frames based on gray-scale intensity, colour and texture have been proposed. On the basis of these measures, one is able to find sharp peaks in plots to determine frame boundaries where the scene is supposed to change. Unfortunately, the identification of correct boundaries is invariably linked to the determination of a correct threshold above which peaks represent genuine scene changes. If the threshold is not optimally set, then a large number of false alarms will result. In this paper we investigate a method of automatically determining the number of scenes in a given video sequence, and therefore automating the process of threshold determination. Our methodology is based on the use of fuzzy clustering of frames that are indexed using texture features. Cluster validity is determined by Davies Bouldin index and a newly proposed temporal validity index. The experimental results show that the proposed method offers a powerful methodology for determining the correct number of scenes and their boundaries in raw video sequences.

1. INTRODUCTION

Successful video segmentation is necessary for most multimedia applications. In order to analyse a video sequence, it is necessary to break it down into meaningful units that are of smaller length and have some semantic coherence. Rui et al. [6] describe a video shot as: "An broken sequence of frames recorded from a single camera. It is the building block of a video. It is a physical entity and is delimited by shot boundaries." Video scenes are further defined as a collection of semantically related and temporally adjacent shots, depicting and conveying a high level concept or story. A range of similarity issues is discussed in [1]. Once video shots are available, higher-level segments can be built. Research into using shots or scene for building higher level information include Scene Transition Graphs, Model based video editing and film theory, Joint analysis of audio, video and text, and the use of Hidden Markov Models for grouping shot structures [10].

Shot boundary detection has been approached by several studies from a variety of perspectives including techniques that are pixel based, statistics based, transform based, feature based and histogram based. It is widely recognised that histogram based and feature based approaches offer the best solution to the problem. Sethi and Patel [9] discuss a number of statistical measures for scene detection. These include OMC (Observer Motion Coherence), NDE (Normalised Difference in Energy), ADSR (Absolute Difference to Sum Ratio Normalised), P (Likelihood), YLR (Yakimovsky Likelihood Ratio), X square (Chi-square), and KS (Kolmogorov-Smirnov Statistic). In addition, other methods based on comparing probability distributions of frames have been used in such research including SHD (Squared Histogram Difference), CME (Cross Entropy Method) [5], KLD (Kullback Liebler Distance), Divergence, Bhattacharya Distance, and HC (Histogram Comparison). In Appendix I we detail these measures.

Statistical measures of calculating shot transitions are based on finding differences between two or three consecutive frames. This task is not easy. In Figure 1 we show a statistic that calculates the dissimilarity between frame pairs across a 20 frame sequence. A threshold (drawn as a straight horizontal line) defines the level above which we consider a peak to represent genuine scene change. Now consider the threshold as 50, and we can see that there are two scene changes at frame numbers 10 and 12. Now consider a threshold of 40, we have 4 scene changes and at a threshold of 30 we get 7 changes. As we lower the threshold, more and more false alarms are introduced. Obviously, it is crucial to know which threshold to set for a given video sequence.

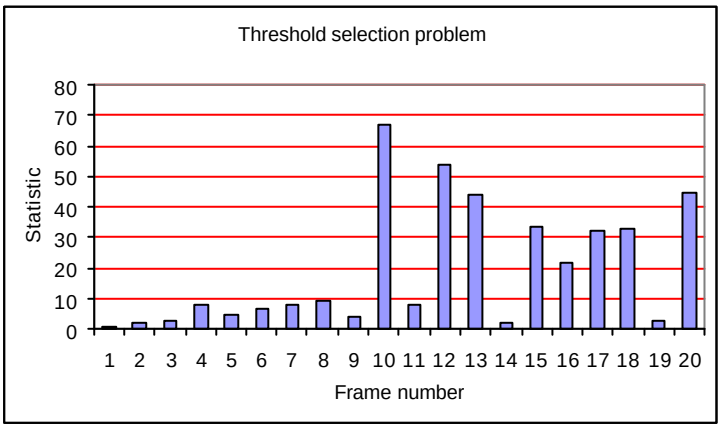


Figure 1. The problem of threshold selection

In this paper we are interested in knowing the number of distinct scenes within a given video. This knowledge is helpful in setting the correct threshold so that genuine scene changes can be identified. In this paper we show how texture based frame clustering can be used to obtain a good estimate of the number of scene present and thus detect appropriate scene changes. In section 2 we detail feature based frame clustering. A temporal validity index is introduced to calculate whether frames that cluster together have some temporal relationship. We detail the use of Davies Bouldin index to determine the valid number of clusters based on inter spatial distances. In section 3 we detail our experimental set-up on a bicycle sequence video data. Section 4 details the results and finally we summarise our finding.

2. FEATURE BASED FRAME CLUSTERING

As discussed earlier, one of the main problems with various measures of frame similarity is that we need an optimal threshold to determine which pair wise frame changes represent a change in scene. It can be reasonably expected that frames that cluster well with each other will come from the same scene. Pixel based clustering is a crude method and feature based methods for this purpose are to be preferred. Clustering can be performed using a range of video features including region shape, colour and texture [7]. One method of clustering frames can be based on the use of global texture. Here we are interested in gray-scale texture (for colour texture features, see [4]). As successive frames in the same scene change only slightly in their features such as their global texture, we expect that if we are to cluster frames on the basis of texture features, frames within the same scene should cluster together. In this paper we use autocorrelation texture features for describing each frame.

An autocorrelation function can be evaluated that measures this coarseness. This function evaluates the linear spatial relationships between primitives. If the primitives are large, the function decreases slowly with increasing distance whereas it decreases rapidly if texture consists of small primitives. However, if the primitives are periodic, then the autocorrelation increases and decreases periodically with distance. The set of autocorrelation coefficients *C* shown below are used as texture features [8]:

$$C_{ff}(p,q) = \frac{MN}{(M-p)(N-q)} \frac{\sum_{i=1}^{M-p} \sum_{j=1}^{N-q} f(i,j) f(i+p, j+q)}{\sum_{i=1}^{M} \sum_{j=1}^{N} f^2(i,j)}$$

, where p, q is the positional difference in the i, j direction, and M, N are image dimensions. We obtain a total of 99 texture features for each image frame in a video sequence. Principal components analysis is next used to extract lower dimensional features that retain the overall variability in feature data. A total of five principal components are extracted.

The feature data is next subjected to a clustering process using fuzzy c-means clustering[2]. The number of clusters to be derived from raw data needs to be specified in advance. This initialisation is necessary for all data vectors to be assigned to a unique cluster. At the end of the clustering process, each image frame is assigned to a unique cluster. There are two important questions at this stage. First, whether we can use the clusters themselves as defining scenes. Second, how to determine the correct number of clusters as their number defines some understanding of how many scenes there possibly are. On the first issue, we find that clusters themselves can not be defined as scenes. Even though a given cluster will in most cases contain contiguous frames, it is necessarily not the case. So their number is more important than their composition. The only point of relevance here is that clusters must have some temporal validity, i.e. ideally, good clustering should find cluster members that are in some temporal sequence rather than out of sequence. For example, a sequence of frames {2,3,4,5,6} is a sequence showing good temporal validity of the clustering process rather than {2,5,9,11,15}. Hence we define a temporal validity index (*tvi*) for a given cluster j whose frames are sorted in ascending order of their number in the original sequence $\left(z_1, z_2, ..., z_{N_j}\right)$ as follows:

$$tvi(j) = \frac{N_j - 1}{\sum_{i=0}^{N_j - 1} \left(1 + \log(z_{i+1} - z_i)\right)}.$$

The overall temporal validity index for a given set of clusters C is given by: $tvi = \sum_{j=1}^{C} tvi(j)$. On the second issue of the number of clusters, we use the Davies Bouldin index [11] which aims to maximise the distances between clusters but at the same time minimising the distances of points in a cluster from their respective centroids. Consider a total of C clusters. We can define inter-cluster distance $S(Q_k)$ and between cluster distances $d(Q_k, Q_l)$. Now considering the fact that samples $x_i, x_{i'} \in Q_k$, $i \neq i'$, $x_j \in Q_l$, $k \neq l$, and N_k is the number of samples in cluster Q_k, the cluster centroids are defined

as: $c_k = 1/N_k \sum_{x_i \in Q_k} x_i$. The inter-cluster distance is given by $S_c = \frac{\sum_i \| x_i - c_k \|}{N_k}$, and the between cluster distance is given by: $d_{ce} = \| c_k - c_l \|$. Davies Bouldin index aims to minimise the following function:

$\frac{1}{C} \sum_{k=1}^{C} \max_{l \neq k} \left\{ \frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_l)} \right\}$. By plotting the index value for different number of clusters, we can determine which one gives the lowest index value.

It should be noted here that our approach is not the only logical method of finding scene changes by clustering. It is also possible to cluster the results of the similarity metrics as a two class clustering problem: scene change and no scene change [3]. This will identify peaks that represent scene changes and no scene changes.

3. EXPERIMENTAL DETAILS

In our experiments we have chosen a bicycle video sequence. This sequence has a total of 236 frames. The genuine scene changes occur at frames 31 and 63. This is shown in Figure 2 below.

Video – Frame 31

Video –Frame 32

Video – Frame 63

Video –Frame 64

Figure 2. Video scene changes at frame numbers 31 and 63.

In order to detect changes between different consecutive frames, we first calculate the various similarity measures shown in Appendix I. Video data is used in uncompressed form and only gray-scale image intensity information is used for our analysis. Next, we extract autocorrelation texture features from each of the frames and extract five principal components. The fuzzy-c-means algorithm is used to cluster texture data into C clusters where C ranges between 2 and 9. The clustering process assigns each frame to a unique cluster with some measure of membership. For each value of C, we calculate Davies Bouldin index and plot the index value. For each cluster, we also calculate the temporal validity index. On the basis of these measures, we select an optimal number of clusters.

Considering that $C_{optimal}$ is the correct number of clusters or scenes, we lower the threshold for each similarity measure till ($C_{optimal} - 1$) peaks are identified. The frame numbers corresponding to these peaks indicate scene changes.

4. RESULTS

In Figure 3 we show the frame similarity measurements. As expected, there are sharp peaks at frames 31 and 63. However, different statistics are measured on different scales and require their own thresholds. Lowering the thresholds will increase false alarms more radically on some measures than others. For example, plots with multiple peaks are more likely to raise false alarms as the threshold is lowered.

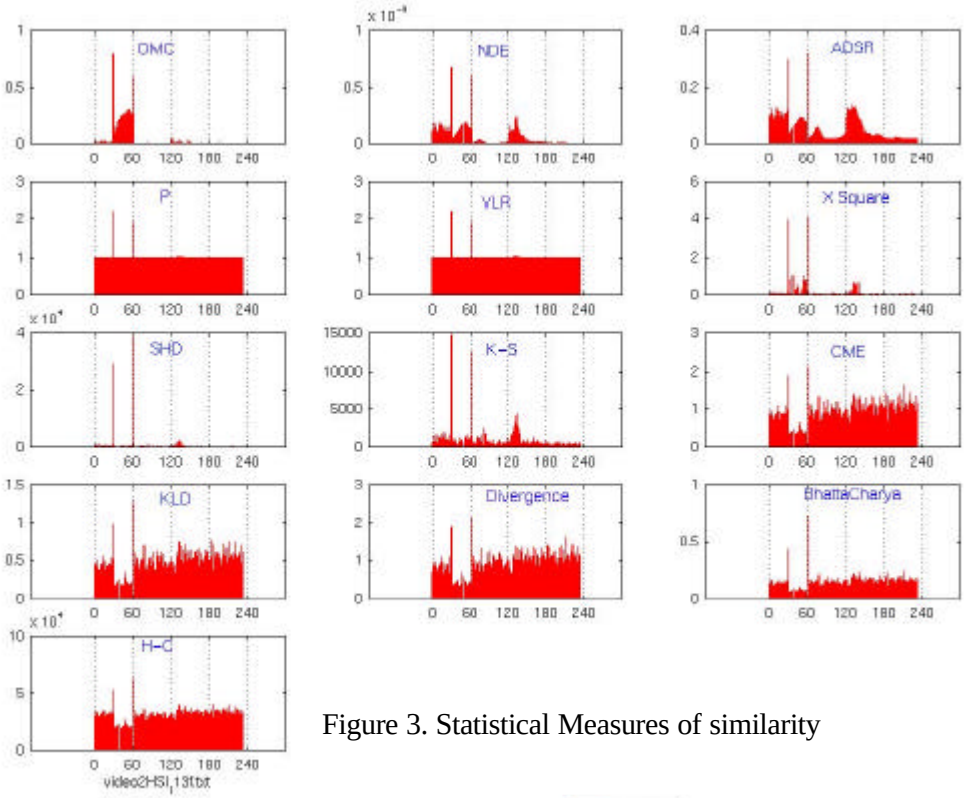


Figure 3. Statistical Measures of similarity

We therefore need to ascertain how many different scenes there are in the video. This is determined by texture based clustering of frames. As we increase the parameter for number of clusters desired, larger clusters are split into smaller ones without much migration of frames across the cluster space. We show this in Figures 4(a-h) where the first two principal components are plotted. We next calculate the Davies Bouldin index for these clusters. This is shown in Figure 5.

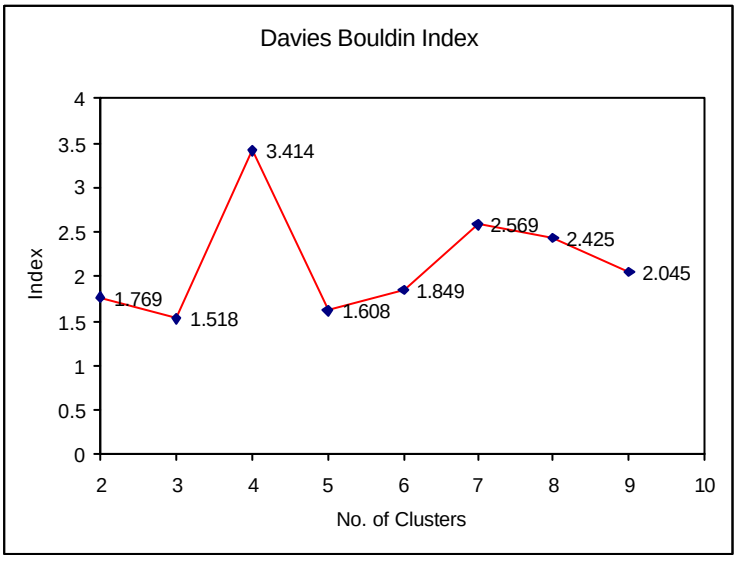
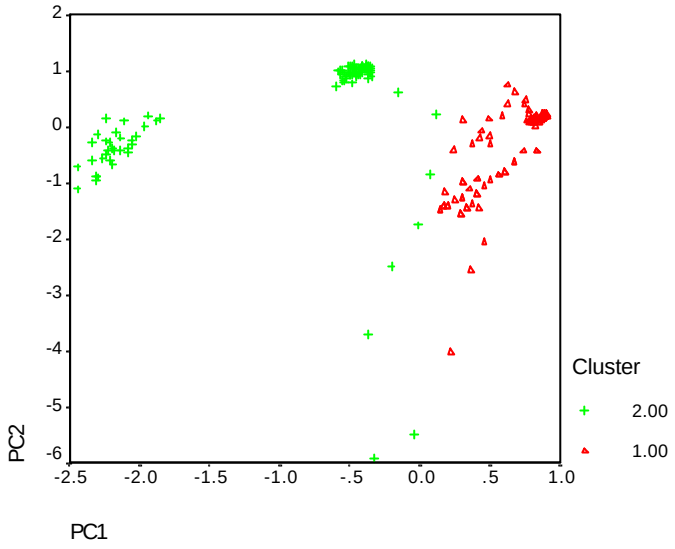
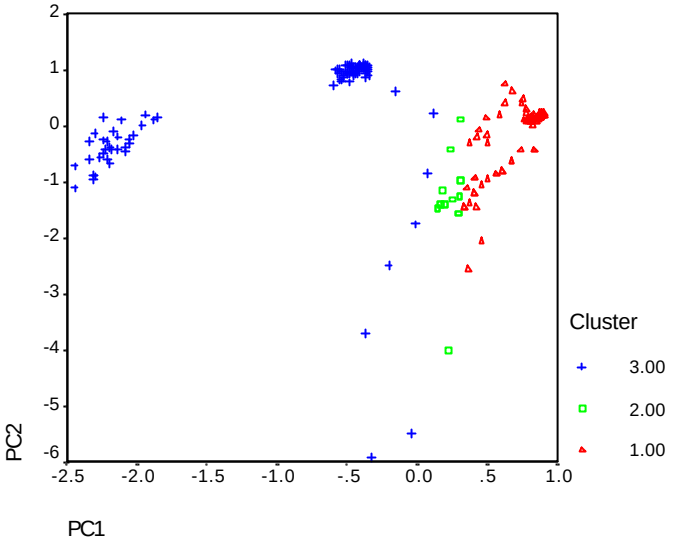


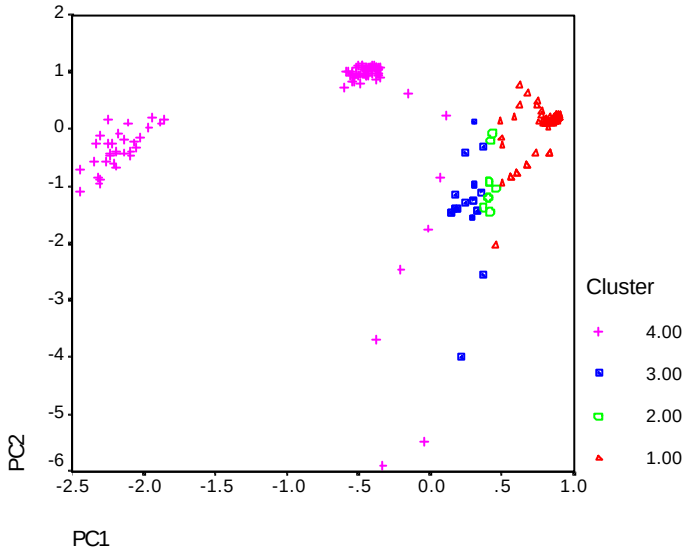
Figure 5. Davies Bouldin index for the bicycle video clusters


PC2
PC1
Cluster
2.00
1.00

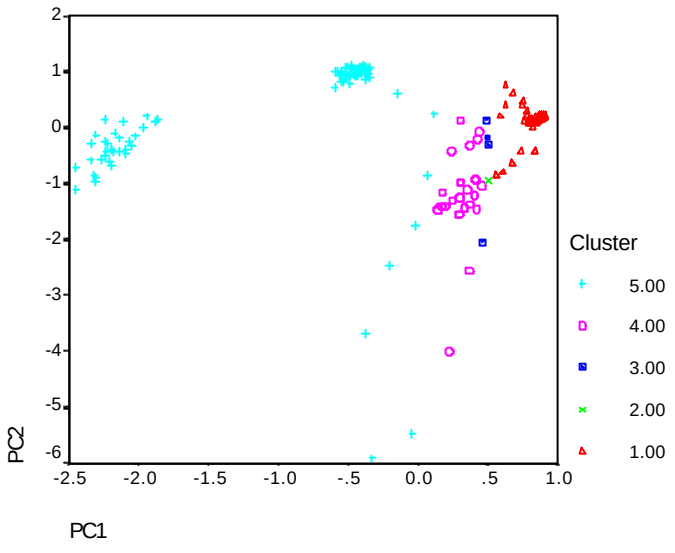
(a)


PC2
PC1
Cluster
3.00
2.00
1.00

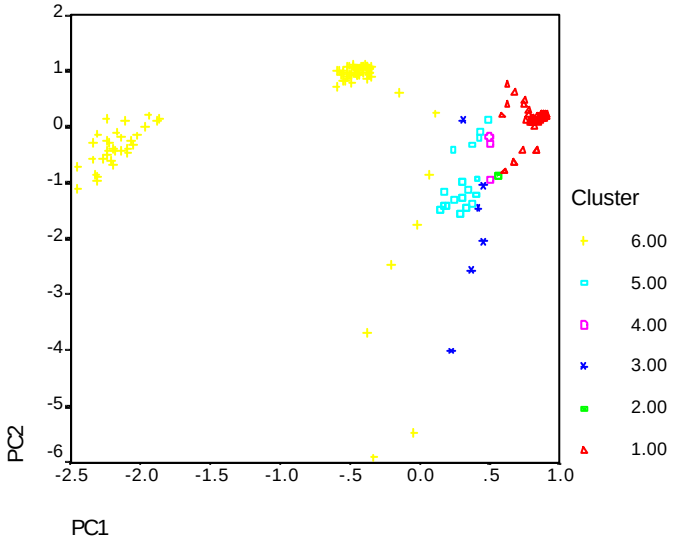
(b)


PC2
PC1
Cluster
4.00
3.00
2.00
1.00

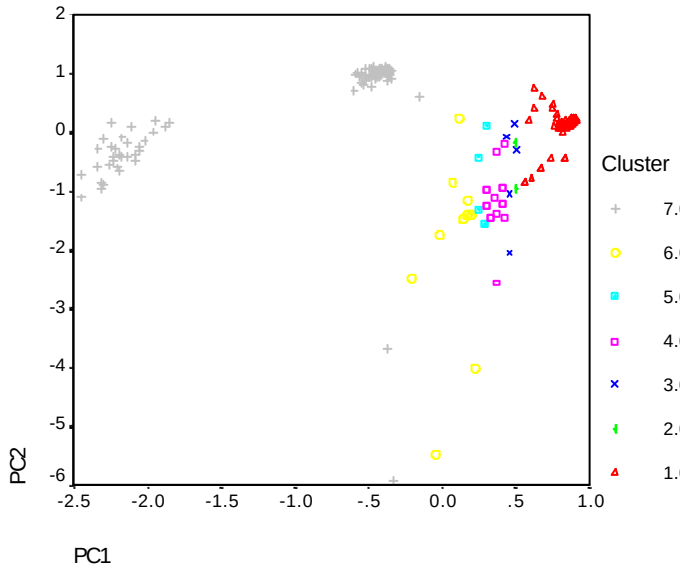
(c)


PC2
PC1
Cluster
5.00
4.00
3.00
2.00
1.00

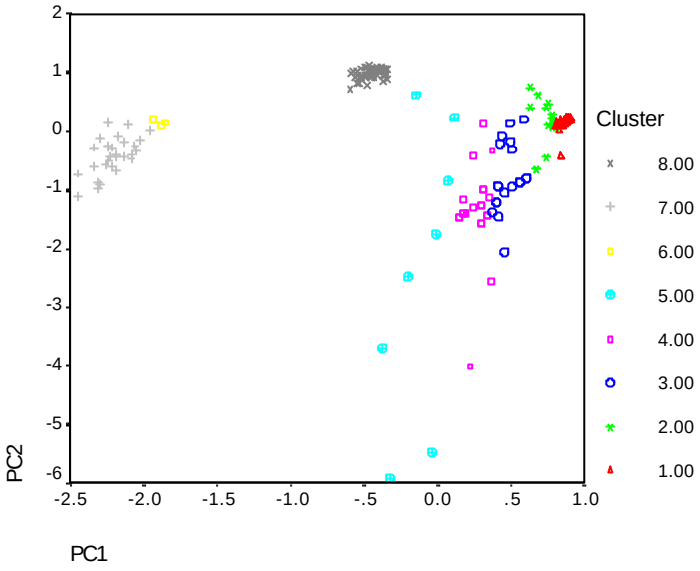
(d)


PC2
PC1
Cluster
6.00
5.00
4.00
3.00
2.00
1.00

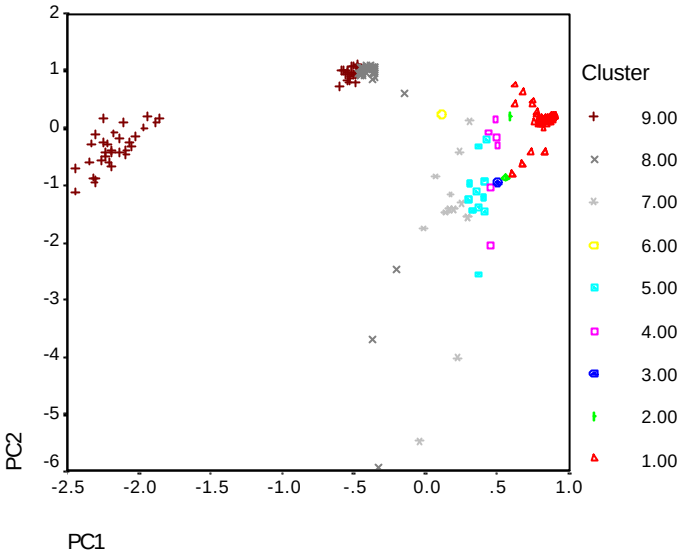
(e)


PC2
PC1
Cluster

(f)



(e)



(h)

Figure 4(a-h). The clusters generated by the fuzzy c-means clustering staring with 2 clusters in 4(a) to 9 clusters in 4(h)

As the best clustering aims to minimise the index, we find that 3 clusters is the best for this video sequence which identifies three separate sequences. For each clustering, we also calculate the temporal validity index. This is shown in Figure 6 below.

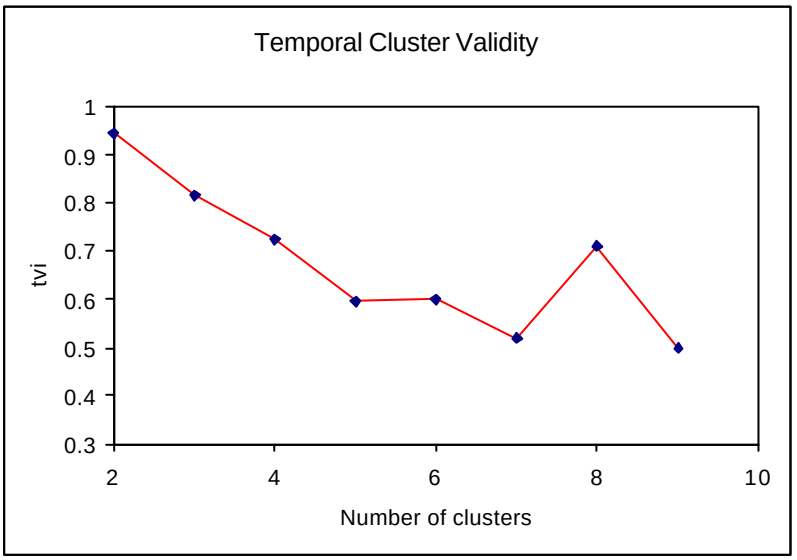


Figure 6. Temporal Validity Index (tvi) plot for different number of clusters

From Figure 6 we can see that as more clusters are created, the index decreases showing that frames in these newly created clusters are out of sequence. The highest value of the index is 1 which is to be preferred. Hence, some form of compromise between the generation of clusters based on the inter-spatial distance between their features (defined by the Davies Bouldin index) and the temporal validity of clusters (defined by the *tvi* index) is needed. This is however context dependent depending on whether we prefer clusters that have frames in a temporal sequence over those without in a given application.

On the basis of the above results, we can safely conclude that our video sequence consists of three scenes for which we need two boundaries of separation. The identification of these two highest peaks from various graphs of similarity measures is now relatively easy and we can statistically confirm that the scene changes occur at frames 31 and 63 which is also visually confirmed in Figure 2.

5. CONCLUSIONS

In this paper we have shown a novel method of identifying scene changes in video sequences. The clusters generated themselves do not define scenes as in some other studies. The reason for this is that such clusters contain frames out of sequence and we would prefer scenes where frames are in sequence. Hence, the clustering process only aids rather than defines the scenes. The optimal number of clusters are used to find the correct thresholds for analysing measures of similarity from where the true peaks representing scene changes can be identified. In our analysis we have only used one texture extraction method and one measure of cluster validity. Obviously, this can be extended to use a variety of methods and a consensus from these can be used. We have also defined in this paper a temporal validity index that weights the gaps between frames by the logarithm of the difference. We expect that our methodology goes some way to solving the fundamental threshold problem for a range of similarity measures. Our further experiments are now aimed at ensuring that the methodology can be calibrated to more difficult video sequences.

REFERENCES

1. P. Aigrain, H.J. Zhang and D. Petkovic, Content-based representation and retrieval of visual media: a state of the art review, Multimedia Tools and Applications, vol. 3, pp.179-202, 1996.
2. R.O. Duda, P.E. Hart and D.G. Stork, Pattern classification, John Wiley, 2001.
3. B. Günsel, A.M. Ferman and A.M. Tekalp, Temporal video segmentation using unsupervised clustering and semantic object tracking, Electronic Imaging, vol. 7, no. 3, pp. 592-604, 1998.
4. I. Ide, R. Hamada, S. Sakai and H. Tanaka, Relating graphical features with concept classes for automatic news video indexing, IJCAI-99 Workshop on Intelligent Information, Stockholm, Sewden, 1999.
5. S.H. Kim and R.H. Park, A novel approach to scene change detection using cross entropy, Proc. ICIP Conference, Vancouver, 2000.
6. Y. Rui, T.S. Huang, and S. Mehrotra, Constructing Table-of Contents for videos, ACM Journal of Multimedia Systems, vol. 7, no. 5, pp. 359-368, 1999.
7. Y. Rui, T.S. Huang and S.F. Chang, Image retrieval: past, present and future, Journal of Visual Communication and Image Representation, vol. 10, pp. 1-23, 1999.
8. N. Sebe and M.S. Lew, Texture features for content-based retrieval, in Principles of Visual Information Retrieval, M.S. Lew (ed.), pp. 51-82, Springer, 2001.
9. I.K. Sethi and N. Patel, A statistical approach to scene change detection, SPIE Conference Proceedings on Storage and Retrieval for Image and Video Databases III, vol. 2420, pp. 329-339, 1995.
10. E. Veneau, R. Ronfard and P. Bouthemy, From video shot clustering to sequence segmentation, Proc. ICPR Conference, vol. 4, pp. 254-257, 2000.
11. J. Vesanto and E. Alhoniemi, Clustering of the self-organising Map, IEEE Transactions on Neural Networks, vol. 11, no. 3, pp. 586-600, 2000.

APPENDIX I

Statistical Measures of Video Shot Segmentation

OMC (Observer Motion Coherence)
Letting S_{ij} and S_{jk} as measures of similarities between consecutive frame pairs (F_i, F_j) and (F_j, F_k) respectively, then $OMC(i, j, k) = \left| \frac{(S_{ij} - S_{jk})}{(S_{ij} + S_{jk})} \right|$ which returns a value close to one if there is scene change in the three frames under consideration, and a value close to zero otherwise.

NDE (Normalised Difference in Energy)
Let $F_i(x, y)$ be the intensity of image location (x, y) in the current frame i, and $F_j(x, y)$ is the image intensity at location (x, y) in the previous frame j, then NDE is defined as follows:

$$NDE = \frac{\sum_{x,y}[F_i(x,y) - F_j(x,y)]^2}{\left[\sum_{x,y} F_i^{\,2}(x,y)\right]\left[\sum_{x,y} F_j^{\,2}(x,y)\right]}$$

ADSR (Absolute Difference to Sum Ratio Normalised)
Let $F_i(x, y)$ be the intensity of image location (x, y) in the current frame i, and $F_j(x, y)$ is the image intensity at location (x, y) in the previous frame j, then ADSR is defined as follows:

$$ADSR = \frac{\sum_{x,y} \left| F_i(x,y) - F_j(x,y) \right|}{\sum_{x,y} \left| F_i(x,y) + F_j(x,y) \right|}$$

P (Likelihood)
An image is subdivided into a set of blocks and a block-wise likelihood ratio is computed, between the (i)th and (j)th frames.

$$P = \frac{\left[\left(\frac{s_i^{\,2} + s_j^{\,2}}{2} \right) + \left(\frac{m_i - m_j}{2} \right)^2 \right]^2}{s_i^{\,2} \cdot s_j^{\,2}}$$ where m and s represent the mean and standard deviations.

YLR (Yakimovsky Likelihood Ratio)
This measure is defined by

$$YLR = \left(\frac{s_0^{\,2}}{s_{i-1}^{\,2}} \right) \left(\frac{s_0^{\,2}}{s_i^{\,2}} \right)$$

where $s_{i-1}^{\,2}$ and $s_i^{\,2}$ represent the variances of the pixel intensity values of the previous and current frames, and $s_0^{\,2}$ denotes the variance of the pooled data from both the frames.

c^2 (Chi-square) Test

The chi-square test is the most accepted test for comparing two binned distributions. $c^2 = \sum_{k=0}^{G-1} \frac{(H_{i+1}(k) - H_i(k))^2}{(H_{i+1}(k) + H_i(k))^2}$, where $H_i(k)$ is the frequency of the gray level k occurring in frame i.

KS (Kolmogorov-Smirnov Statistic)

Letting CH_i and CH_{i+1} represent the cumulative histogram sum up to bin k for the previous and current frame respectively, the statistic is defined as:

$$KS = \max_k |CH_{i+1}(k) - CH_i(k)|$$

SHD (Squared Histogram Difference)

This is given for pair-wise frame comparison as follows:

$$SHD = \sum_{k=0}^{G-1} \frac{(H_{i+1}(k) - H_i(k))^2}{(H_{i+1}(k) + H_i(k))^2}$$

HC (Histogram Comparison**)**

$HC = \sum_{k=0}^{G-1} |H_{i+1}(k) - H_i(k)|$ for grey level

CME (Cross Entropy Method)

Let p be the probability density function (pdf). Let $p_{i+1}(x)$ and $p_i(x)$ be the probability of intensity level x in $(i+1)$th frame and i-th frame respectively. Then CME measure is defined as:

$$CME(i, i+1) = \int p_{i+1}(x) \log \frac{p_{i+1}(x)}{p_i(x)} dx$$

Divergence

This is defined as:

$$Divergence(i, i+1) = \int p_i(x) \log \frac{p_i(x)}{p_{i+1}(x)} dx + \int p_{i+1}(x) \log \frac{p_{i+1}(x)}{p_i(x)} dx$$

KLD (Kullback Liebler Distance)

$$KLD(i, i+1) = \int p_{i+1}(x) \log \frac{p_{i+1}(x)}{p_i(x)} dx + \int (1 - p_{i+1}(x)) \log \frac{(1 - p_{i+1}(x))}{(1 - p_i(x))} dx$$

BD (Bhattacharya Distance)

$$BD = -\log \int (p_{i+1}(x)\, p_i(x))^{\frac{1}{2}} dx$$