

# Video Indexing and Retrieval in Compressed Domain Using Fuzzy-Categorization

Hui Fang, Rami Qahwaji, and Jianmin Jiang

EIMC Dept., University of Bradford  
{h.fang, r.s.r.qahwaji, j.jiang1}@bradford.ac.uk

**Abstract.** There has been an increased interest in video indexing and retrieval in recent years. In this work, indexing and retrieval system of the visual contents is based on feature extracted from the compressed domain. Direct processing of the compressed domain spares the decoding time, which is extremely important when indexing large number of multimedia archives. A fuzzy-categorizing structure is designed in this paper to improve the retrieval performance. In our experiment, a database that consists of basketball videos has been constructed for our study. This database includes three categories: full-court match, penalty and close-up. First, spatial and temporal feature extraction is applied to train the fuzzy membership functions using the minimum entropy optimal algorithm. Then, the max composition operation is used to generate a new fuzzy feature to represent the content of the shots. Finally, the fuzzy-based representation becomes the indexing feature for the content-based video retrieval system. The experimental results show that the proposal algorithm is quite promising for semantic-based video retrieval.

## 1 Introduction

Video indexing and retrieval is becoming an important application because of the increasing demand for such systems that could provide effective organization, search and display of multimedia contents. This is caused by the tremendous growth of the multimedia archives, particularly in the digital video database [2] [4] [5] [6] [9]. In order to analyze the content of the videos, the meaningful information needs to be extracted. However, any single low-level feature is not adequate to provide full and efficient representation for videos. Therefore, a hyper feature extraction mechanism is important for video indexing and retrieval systems.

Systems based on fuzzy logic usually require fewer variables, use a linguistic rather than a numerical description, and can relate output to input by analyzing and understanding the data as a human being. Fuzzy logic has been used for video processing in recent years. In [14], fuzzy logic and fuzzy rules were defined to provide semantic interpretation for video contents. However, in [14] fuzzy logic was used for obtaining the rules but the membership functions were set manually. In this paper, an automated technique for determining the fuzzy membership functions is presented using the minimum entropy algorithm.

The traditional video indexing and retrieval process has four steps: temporal segmentation, key frames selection, feature extraction and classification. In this paper,

we assume that the shot boundaries have been detected and all the videos are organized in shots. The key-frame selection algorithm [1] [2] extracts some frames to represent the visual content in every shot. Following these two steps, two types of features, the spatial features and the temporal features, can be extracted separately from the shots and the key-frames. The spatial features are mainly visual features such as color, texture and shape which are extracted from key-frames by image analysis techniques. In [8], image database retrieval is carried out using the mean and variance for the local blocks that were extracted from DCT domain to create a histogram. In [11], the authors use the average color feature in the Luv space as a spatial feature to match the frames. The motion features represent the camera and objects movements and many motion descriptors are designed. Motion activity [3], which is provided by The MPEG-7 standard, shows the intensity of the activity. The magnitude of block-based motion vectors is directly obtained from MPEG compressed domain and its standard deviation is classified into five bins. The Pixel Change Ratio Map (PCRM) [4], which is motivated by the frequency of the pixel intensity changes, is another temporal feature characterizing motion content in a video shot. If the objects in a shot move faster, the pixel intensity changes more. A tensor histogram extracted directly from temporal slice [5] is another good motion descriptor to index and organize the motion content of video shots. In this method, the tensor histogram is compared with two more features, Gabor feature and Co-Occurrence Matrix to prove it is an efficient feature in the compressed domain. A descriptor using block-based motion vector and PCA is given in [6]. To compare the efficiency of these motion features, in [7] a framework to measure the intensity of motion activity of video shots is built to test the performance of some temporal descriptors and the results show that the MPEG-7 motion activity descriptor is one of the best performers. When obtaining all the features for retrieval, the final step is matching the features between the query shots and database shots. [9] [10] use a total distances which is obtained based on the weighted distances of various features while [5] use a hierarchical structure to match these two types of features sequentially. Using these steps, the retrieval results are ranked by the similarity distances between the query example and the descriptors in the video database.

Although these traditional technologies have got some reasonably good performance, they still have some drawbacks: (1) some features are high-dimensional features, which increase the computational complexity of the whole process and affect the retrieval speed; (2) some features are classified into different semantic classes by crisp threshold which makes the system inflexible and damages the performance; (3) a mechanism is required to integrate multi-features into a retrieval system. As illustrated before, fuzzy representation is used because it gives a more compact description for improving the retrieving performance [12]. Furthermore, it enables a human-based interpretation for the discriminative features. In addition, a new feature can be generated by combining different features in different fuzzy sets through the fuzzy composition relations. In this paper, a fuzzy-categorizing video indexing and retrieval paradigm is constructed and a new feature for comparing the similarity between the shots is exploited to improve the performance of the video retrieval system.

This paper is organized as follows: The proposed algorithm is proposed in section 2. Experimental results on a basketball video database are shown in section 3. Finally, the concluding remarks are provided in section 4.

## 2 Algorithm Design

### 2.1 Paradigm of the Video Indexing and Retrieval

Video Database and multimedia library indexing and retrieval system plays an important role in the multimedia application field. Despite the intensive research efforts, the development of low-level feature-based video retrieval algorithm is still in its early stage. Linking the low level features to high level semantics remains an open problem in computer vision [14]. In this paper, an inter-layer fuzzy feature is designed to bridge the gap between the huge amount of data representing low-level features and the meaningful content of videos. The low-level perceptual features are trajectory to a semantic fuzzy domain and the retrieval system retrieve the content of videos based on the integrated fuzzy features. In other words, the low-level features are grouped into some meaningful categories by the fuzzy logic and then the fuzzy values of those features are integrated to index the shots in the video database.

At present, video compression standards such as MPEG2, MPEG4, H263 and H264 have been designed to reduce the storage and transmission requirements. Therefore, the video indexing and retrieval in the compressed domain is a promising field. In this paper, temporal features and spatial features that are extracted from the compressed domain are employed in the fuzzy system for video shot-based retrieval. The details of these two features are given below:

#### 2.1.1 Statistical Spatial Variance Feature

Given a shot with  $K$  I-frames (which are the intra-coded frames) and  $M \times N$  blocks with DCT coefficients in each frame, the variance of each block, based on [8], is given in equation (1),

$$\text{var} = \frac{1}{64} \sum_{u=0}^7 \sum_{v=0}^7 C^2(u, v), \quad (u, v) \neq (0, 0) \quad (1)$$

where  $C(u, v)$  is the AC coefficient in this block. According to Equation (1), the spatial average variance for each frame is shown in equation (2),

$$AV = \frac{1}{M \times N} \sum_{m=0}^M \sum_{n=0}^N \text{var}_{mn} \quad (2)$$

where  $AV$  represents the average variance for each I-frame and  $\text{var}$  represents the same variance in equation (1). The spatial average variance for the whole shot is given as follow:

$$SF = \frac{1}{K} \sum_{k=0}^K AV_k \quad (3)$$

where  $K$  is the total number of I-frames in the shot. Variance is a key feature in video and image retrieval because it reveals the texture on the spatial image or frame and the variance is similar for the same semantic content. For example, the bird-view of the crowds has a large variance while the close-up of some players has a relative smaller variance value.

**2.1.2 Statistical Temporal Deviation Feature**

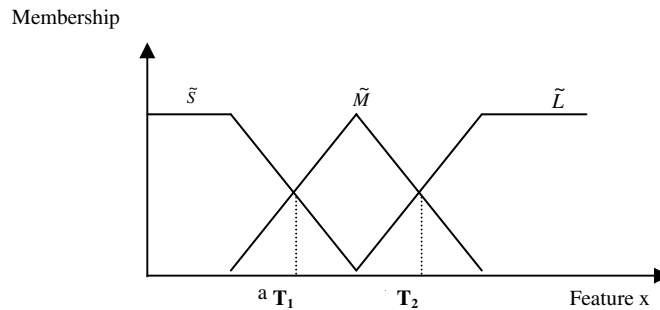
This feature is derived from the motion activity feature in MPEG-7 [3]. Instead of quantizing the motion magnitude deviation into 5-bins histogram, the average of the motion magnitude deviation of the shot is considered to be the motion feature. Motion features are important for video processing because the temporal continuity of visual cues is unique in a video stream. This feature does not exist for static images.

When the features are extracted, they are transformed into the fuzzy domain to relate them to some semantic concepts. In this paper, three concepts, which include full-court match, penalty and close-up, are used to train the fuzzy membership functions. The details are shown in the next subsection. Then the max composition operation is used to generate a 9-bins fuzzy feature to represent the semantic meaning for every shot. Then the correlation between the query shot and the shots in the database is calculated based on this semantic fuzzy feature.

We believe that there are three advantages for using this feature: it is (a) a novel feature linking the low-level features to the semantic categories; (b) a novel feature to represent the shot compactly and efficiently; (c) a novel feature to integrate the temporal feature and the spatial feature.

**2.2 The Generation of Fuzzy Membership Functions**

At the first step, fuzzy membership functions are required in order to fuzzify the features into the fuzzy sets. A common fuzzy membership function is shown in Figure-1,



**Fig. 1.** Illustration of the fuzzy membership function

where three fuzzy sets S, M and L are shown to represent the fuzzy sets in the feature domain. The corresponding membership functions are defined as follows:

$$\begin{aligned}
 M(x)_L &= \begin{cases} 1 & \text{if } x > c \\ \frac{x-b}{c-b} & \text{if } b < x < c \end{cases} \\
 M(x)_M &= \begin{cases} \frac{c-x}{c-b} & \text{if } b < x < c \\ \frac{x-a}{b-a} & \text{if } a < x < b \end{cases}
 \end{aligned}
 \tag{4}$$

$$M(x)_s = \begin{cases} 1 & \text{if } x < a \\ \frac{b-x}{b-a} & \text{if } a < x < b \end{cases}$$

where  $x$  represents a feature, and the three parameters  $a$ ,  $b$ , and  $c$ , represent the boundaries of the three fuzzy sets as shown in Figure 1.

In this paper, the minimum entropy optimization threshold algorithm [13] is used to train the membership function in order to fuzzify the motion features and the spatial features into the fuzzy sets. As shown in figure-1, two Thresholds  $T_1$  and  $T_2$  are trained by getting the minimum entropy value to separate the three classes. For  $T_1$  and  $T_2$ , the membership values  $M(x)$  of the classes are set to 0.5 [13]. The entropy of the threshold dividing two fuzzy sets can be calculated using the following equations [13]:

$$\begin{aligned} S(x) &= p(x)S_p(x) + q(x)S_q(x) \\ S_p(x) &= -[p_1(x)\ln p_1(x) + p_2(x)\ln p_2(x)] \\ S_q(x) &= -[q_1(x)\ln q_1(x) + q_2(x)\ln q_2(x)] \end{aligned} \tag{5}$$

where  $p_1(x)$  is the probability that class 1 samples are located in region  $p$ ,  $p_2(x)$  is the probability that class 2 samples are located in region  $p$ ,  $q_1(x)$  is the probability that class 1 samples are located in region  $q$ , and finally  $q_2(x)$  is the probability that class 2 samples are located in region  $q$ . The percentage of total number of samples belonging to region  $p$  in all samples is referred to as  $p(x)$ , while the percentage of total number of samples belonging to region  $q$  in all samples is referred to as  $q(x)$ . In this work, regions  $p$  and  $q$  are divided by thresholds  $T_1$  or  $T_2$  as shown in Figure 2.

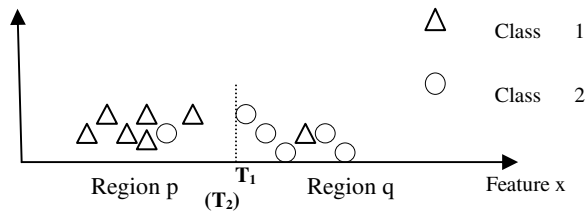


Fig. 2. Illustration of the entropy equation

When finding the minimum entropy thresholds by training the dataset and the parameters are obtained as follow:

For the motion feature:

$T_1 = 2.5, T_2 = 9.2, a = -0.85, b = 5.85, c = 12.55;$

For the variance feature:

$T_1 = 61, T_2 = 87, a = 48, b = 74, c = 100;$

Then, two membership function sets are obtained to fuzzify the two features into the fuzzy domain.

### 2.3 The Generation of Fuzzy Feature and Similarity Measure

After the previous step, each feature is fuzzified into three fuzzy sets. In order to keep the possibility of the principle elements in the fuzzy sets, the max composition operation [13] is used to calculate the possibility that the two features belong to the specific fuzzy sets. Then a nine-bin feature is generated to represent the shot in a semantic layer. At last, the Cosine Amplitude is employed for measuring the similarity between the shots, as shown in equation (6).

$$S_{ij} = \frac{\left| \sum_{k=1}^9 f_{ik} f_{jk} \right|}{\sqrt{\left( \sum_{k=1}^9 f_{ik}^2 \right) \left( \sum_{k=1}^9 f_{jk}^2 \right)}} \quad (6)$$

where  $f_{ik}$  represents the fuzzy feature of the query shot and  $f_{jk}$  represents the fuzzy feature of each shot in the database. The larger  $S_{ij}$  means that the query shot and the retrieval shot are more similar.

## 3 Experimental Results

In this section, the fuzzy system described in the previous section is applied to a basketball database. The database consists of 440 video sequences taken from some NBA basketball matches. This database includes 210 full-court match shots, 70 penalty shots and 170 close-up shots. The duration span of the video shots varies is from 4s to 73s and the frames rate is 29.97frames/sec.



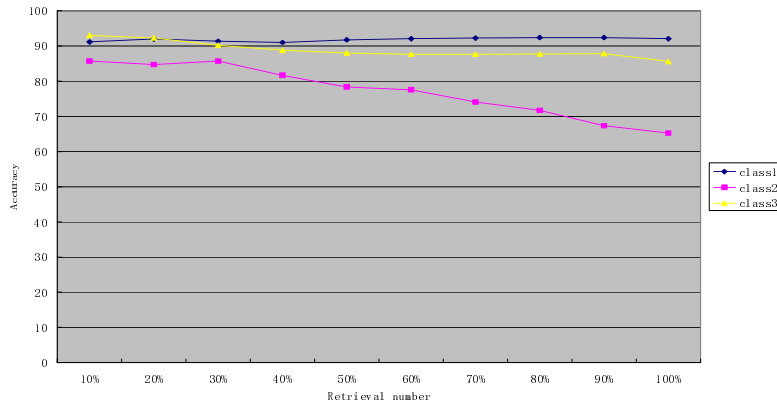
Fig. 3. Examples of video retrieval results for the top 4 best matching shots

The database is used to train the membership functions which have been illustrated in section 2 and the 9-bins fuzzy features are used to index the database. Based on the Jack-Knife technique, we use 90% of this dataset to train the membership function and the remaining 10% are used to test the retrieval performance. Figure 3 shows some examples for our video retrieval experiments. The first frame is exploited to represent the shot.

To further prove the efficiency of our algorithm, the precision of retrieval is considered. The average accuracy of all the retrieval results based on the return number is shown in Table-1, we use percentage as a unit because the numbers of the class samples are different. At the same time, the accuracy for different classes is drawn in figure-4. The shots in class 1 are the full-court match shots, the shots in class 2 are the penalty shots and the shots in class 3 are the close-up shots.

**Table 1.** Retrieval Accuracy according to the num percentage

Retrieval (%)	10	20	30	40	50
Accuracy	93.3	93.0	92.1	90.8	90.3
Retrieval(%)	60	70	80	90	100
Accuracy	90.2	89.8	89.5	88.8	87.4



**Fig. 4.** The Accuracy of the retrieval results

## 4 Conclusion

A new fuzzy feature for retrieving the visual semantic content is presented in this paper. By analyzing the visual content in a basketball video database, three classes, which include full-court match, penalty and close-up, are used to train the fuzzy membership functions of a temporal feature and a spatial feature through the minimum entropy optimal algorithm. Then, the max composition operation is used to generate a 9-bins fuzzy feature. Finally, this representation is applied for the content-based video retrieval system. Experimental results show that this algorithm is a promising solution

for the content based video retrieval system and has a computational efficiency due to the compact feature and indexing process in the compressed domain.

In the near future, we would like to extend this application to provide fuzzy representation for other videos representing different real-life events based on the multi-dimensional feature extraction techniques. And neural network will be used to replace entropy in calculating fuzzy membership functions because of its suitability for higher dimensional features.

## Acknowledgment

The authors wish to acknowledge the financial support from the EU IST Framework-6 programme under the IP project: Live staging of media events (IST-04-027312).

## References

1. A. Hanjalic and H.J. Zhang: An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis. *IEEE Transaction on Circuit and Systems for Video Technology*, Vol.9, No.8, December 1999;
2. A.D. Doulamis and N.D. Doulamis: Optimal Content-based Video Decomposition for Interactive Video Navigation. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.14, No.6, June 2004;
3. B.S. Manjunath, P. Salembier, T. Sikora: Introduction to MPEG-7: Multimedia content description interface. John Wiley& Sons, LTD, ISBN 0471486787, 2002;
4. H. Yi, D. Rajan, L.T. Chia: A new motion histogram to index motion content in video segments. *Pattern Recognition Letters* 26 (2005) 1221-1231;
5. C.W. Ngo, T.C. Pong, H.J. Zhang: "On clustering and retrieval of video shots through temporal slices analysis. *IEEE Transactions on Multimedia*, Vol.4, No.4, December 2002;
6. E. Sahouria and A. Zakhor: Content analysis of video using principal components. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol 9, No.8, Dcember 1999;
7. K. A. Peker and A. Divakaran: Framework for measurement of the intensity of motion activity of video segments. *J. Vis. Commun. Image R.* 15 (2004) 265-284;
8. G. Feng and J. Jiang: JPEG compressed image retrieval via statistical features. *Pattern Recognition* 36 (2003) 977-985;
9. J. Fan, W.G. Aref, A.K. Elmagarmid, M.S. Hacid, M.S. Marzouk, X. Zhu: MultiView: Multilevel video content representation and retrieval. *Journal of Electronic Imaging* 10(4), 895-908 (October 2001);
10. S.F Chang, W. Chen, H.J. Meng, H. Sundaram and D. Zhong: A Fully Automated Content-Based Video Search Engine Supporting Spatiotemporal Queries. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.8, No.5, September 1998;
11. A. Hanjalic, L. Lagendijk, J. Biemond, and J. Biemond: Automated High-level Movie Segmentation for Advanced Video Retrieval System. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.9, No.4, June 1999, pp580-588;
12. A. Doulamis, N.D. Doulamis, S. D. Kollias: A Fuzzy Video Content Representation for Video Summarization and Content-based Retrieval. *Signal Processing* 80 (2000) 1049-1067;
13. Timothy J. Ross: *Fuzzy Logic with Engineering Applications*. John Wiley & Sons, Ltd, 2004;
14. A. Dorado, J. Calic, E. Izquierdo: A Rule-Based Video Annotation System. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.14, No.5, May 2004;