Generic Video Classification: An Evolutionary Learning based Fuzzy Theoretic Approach

R.S. Jadon
CSE Dept.
IIT Delhi
New Delhi-16
rsjadon@cse.iitd.ernet.in

Santanu Chaudhury
EE Dept.
IIT Delhi
New Delhi-16
santanuc@ee.iitd.ernet.in

K.K. Biswas
CSE Dept.
IIT Delhi
New Delhi-16
kkb@cse.iitd.ernet.in

Abstract

In this paper we propose an evolutionary learning based fuzzy theoretic approach for classifying video sequences into generic categories. This categorization is based on video structure based syntactic features. The features like shot durations, editing style, camera work and shot activity conveys large amount of information about the type of video. The information derived from these features is integrated over a larger time-scale than a shot length time to form fuzzy rules for the extraction of video structure based features. Evolutionary learning paradigm is used to evolve the fuzzy rule based system for generic video characterization. Such a rule-based system yields high representational accuracy of the classes as shown by the experiments conducted on various type of video sequences ranging up to 1 to 3 minutes. Experimental analysis illustrates the effectiveness of our system in offering a novel approach for categorizing the video sequences.

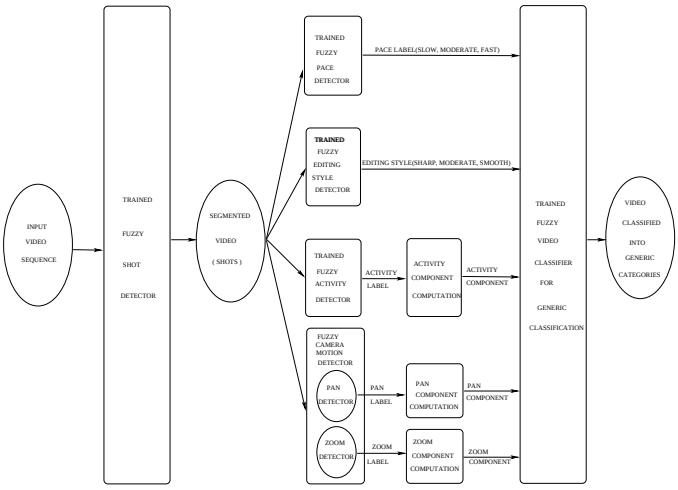

Figure 1: An Overview: Generic Video and Retrieval

1 Introduction

Content processing of visual media is one of the central issues in the present state-of-the-art multimedia technologies. However providing a comprehensive interpretation of video data is a challenging problem because of the considerable heterogeneity in the semantics of different categories of video sequences and inherent variabilities within the categories. In this paper we have proposed an evolutionary learning based fuzzy theoretic scheme for analyzing the video structure. The basic motivation of this analysis is to explore the film theory principles for content based video classification [1]. The overall classification scheme is shown in Fig. 1.

In recent years many approaches have been suggested in the literature for classification and retrieval of image and video data[6, 3, 9, 12, 10, 2]. Most of these schemes are based upon the crisp estimation of syntactic video features. However, due to variations in imaging conditions, object motion in the scene and innovations in film and video technology, crisp estimates do not always provide a robust mechanism for generic video classification. To overcome these problems we have proposed a soft-computing approach using fuzzy logic for the purpose of video characterization. To make the system adaptive a supervised learning based evolutionary algorithm has been used to learn the fuzzy systems. The fuzzy classifiers using fuzzy features provide a mechanism to deal with inherent intra-class variability of the video data.

The rest of the paper is organized as follows: We briefly discuss evolutionary learning paradigm in Section 2. In section 3, the fuzzy syntactic feature extraction is presented. Section 4 describes the generic video classification schemes with illustrations; and Section 5 concludes this paper.

2 Evolutionary Learning of Fuzzy Rule Based System

The fuzzy rule base systems used for feature extraction and classification have been learnt using an evolutionary learning scheme. The learning scheme is proposed in [11]. The main issues involved here are:

- Designing of an encoding scheme for representation of the fuzzy rule based system in terms of chromosomes.
- Generation of initial population.
- Calculation of fitness of each chromosome in the population and selection of chromosomes for reproduction
- Reproduction of new chromosomes using crossover and mutation operation.

The steps of fitness value computation, selection and reproduction is iterated until convergence condition is met. Each chromosome in our system represents a complete fuzzy rule based system (FRBS) encoding membership functions of input and output variables and relevant sets of rules. Integer based encoding is used for this purpose. The fitness function we have chosen accumulates the reward of each individual (i.e. each FRBS) in proportion to the number of correctly classified patterns and the degree of correctness. Through evolutionary learning the most appropriate fuzzy rule based system is evolved.

3 Fuzzy Feature extraction

Editing style and the way in which different types of shots and camera movements are juxtaposed in a film or a video provide reasonably good indication about its genre. A commercial video, typically, has a significant/large component of short shots with large inter-frame variations. The features like shot duration, intra-shot activity and frequency of occurrence of shot transitions can model these characteristics. Camera motion based features show typical patterns in some other types of video. For example, romantic movies, in general, have a large number of close-ups which are followed by zoom-in shots. Sports sequences are associated with predominantly panning/zooming shots.

3.1 Video Editing-Style Based Features

A video sequence is made up of shots. To explore the structure of video we need to understand and analyze the shots. For this purpose temporal video segmentation is a prerequisite. For the purpose of video segmentation, we have developed a fuzzy-theoretic scheme as reported in [7]. This scheme not only detects the shots but also categorizes the shot transitions as abrupt and gradual. Output of the segmentation phase is used for editing style based feature extraction.

3.1.1 Pace Computation

The pace of a sequence is characterized by shot duration and number of shots. In the case of fast paced videos(action/commercial movies) we have a large number of shots of smaller duration, while in slow paced video (news/dialogue movies) shot lengths are large.

In the present approach shot duration is fuzzified as small, moderate and large. We than apply α-cut to these fuzzy sets, with an alpha value of 0.1. The value of alpha is determined experimentally. The resulting fuzzy-sets are represented as α_{small}, $\alpha_{moderate}$ and α_{large}. The cardinality of these sets denoted as $cdn(\alpha_{small})$, $cdn(\alpha_{moderate})$ and $cdn(\alpha_{large})$ are used as one of the features for pace detection. Another feature used is the relative duration of each of these sets computed as follows:

$$nlc_{small} = \frac{\sum_{i=1}^{n} L_i^{small}}{N}$$

where nlc_{small} indicate the normalized total length of the shots of $small$ duration, Li^{small} is the length of i^{th} shot of type $small$, n is the cardinality of fuzzy set $cdn(\alpha_{small})$ and N is the length of the sequence(total number of frames). Similarly we compute the $nlc_{moderate}$ and nlc_{large}. Using these six variables as input we evolve the fuzzy system for characterizing the pace of a video sequences as slow, moderate or fast. The evolved rules have identified useful combinations of cardinality and normalized length for effectively categorizing the perceived pace of the video sequences. These combinations are also not always intuitively obvious. The features extracted from the segmentation results for the purpose of pace computation for a news sequence are shown in Table 1. Here the cardinality of *small* and *moderate* is higher than the cardinality of *large*, but due to high value of normalized-length-component for *large*, it is categorized as a slow-pace sequence (memberships slow:0.700592, moderate:0.415724, fast: 0.078621).

SHOT TYPE	CARDINALITY	NORMALIZED LENGTH COMPONENT
small	07	0.13
moderate	07	0.29
large	05	0.55

Table 1: Fuzzy Shot Durations: Input for Pace Computation

3.1.2 Editing Style Characterization

During video editing process shots are put together using some conventions. Efforts are made to maintain the continuity of the scene keeping in mind the basic principles of film editing [1]. Documentaries and news video, for example, have large number of abrupt transitions, while, in the feature films the gradual transitions are more. Our fuzzy shot detection scheme [7] classifies the shot transitions as fuzzy sets: Abrupt and Gradual. The cardinalities of these sets with alpha cuts(alpha = 0.1), represented as $cdn(\alpha_{abrupt})$ and $cdn(\alpha_{gradual})$, are used as key features for determining the editing style. Another important feature that characterizes editing style is the pattern according to which these transitions are sequenced. This can be measured by temporal duration of such transition patterns. For this purpose, we use normalized length component for abrupt to abrupt transition(nlc_{aa}), gradual to gradual transition(nlc_{gg}) and abrupt to gradual/gradual to abrupt($nlc_{ag/ga}$) transitions computed as follows:

$$nlc_{aa} = \frac{\sum_{i=1}^{n} L_i^{aa}}{N}$$

where, L_i^{aa} is the length of i^{th} shot having both the boundaries of the type abrupt (we consider starting and ending points of the sequences as abrupt transitions), n is the number of such shots and N is the length of the sequence(total number of frames); (nlc_{gg}) and ($nlc_{ag/ga}$) are calculated in the similar manner. These features are used as input to evolve the FRBS for characterizing the editing style as: **sharp, smooth, moderately sharp and moderately smooth**.

Note, that in news sequences most of the transitions are abrupt in continuation therefore these are characterized as a sharp transition sequence with high value, while in advertisement sequence the number of gradual transitions are large but due to discontinuities of their temporal occurrences, the length component of abrupt-to-abrupt transition dominates and therefore these are also characterized as a sharp transition sequence but with low membership values. We have established the ground truth by manual observations and obtained correct characterizations in about 90% cases.

3.2 Motion-Based Features

The motion in video sequences may be due to camera movement, object movement or a combination of the two. In the present work we have analyzed the video structure using motion information. The key feature for this purpose are motion vectors, calculated using Horn and Shunk's [5] optical flow computations. We have estimated the camera motion and shot activity as discussed in the subsequent subsections.

3.2.1 Camera Motion

Information about camera operation is very important for the analysis and classification of video shots, since camera operation often reflects the intentions of the director [4]. There are two important camera operations: panning and zooming. Each of these operations induces a specific pattern in the field of motion vectors from one frame to the next. We have developed a fuzzy theoretic scheme for qualitative characterization of camera motion in a video sequence[8]. Using this scheme we characterize all the shots of a video sequence for panning and zooming motion. We then compute the normalized length component for panning and zooming for a sequence as follows:

$$nlc_{panning} = \frac{\sum_{i=1}^{n} L_i^{panning}}{N}$$

Where n is number of panning shots, $L_i^{panning}$ is the length of i^{th} panning shot and N is the length of the complete sequence.

$$nlc_{zoom} = \frac{\sum_{i=1}^{m} L_i^{zoomin} + \sum_{i=1}^{n} L_i^{zoomout}}{N}$$

Where m is number of zoom-in shots, n is number of zoom-out shots, L_i^{zoomin} is the length of i^{th} zoom-in shot, $L_i^{zoomout}$ is the length of i^{th} zoom-out shot and N is the length of the complete sequence.

$nlc_{panning}$ and nlc_{zoom} are used as the representative features for panning and zooming motion respectively for the complete sequence in the final categorization.

3.2.2 Shot Activity

Shot activity is an important feature for movie classification [12]. Shots perceived as **high activity**(violence shots, football shots, road side scenes) have several objects with rapid and random movements as compared to the shots perceived as **low activity**(television anchor shot, stills, talk shows) having few with small or no movement. We have estimated the overall shot activity called global shot activity and spatial distribution of activity across the frames called local shot activity.

Global Shot Activity Estimation To estimate the global activity the magnitude histogram of optical flow values is used. In high activity shots the motion is distributed among all the bins while in low activity shots it is concentrated among first few bins. We first compute the histogram of the magnitude of the optical flow and then divide it into five unequal partitions called tiles. Intuitively this partitioning is such that the lower order tiles characterizes the low activity while higher order tiles characterizes the high activity. The

Sequence	sharp	smooth	mod-sharp	mod-smooth
News	0.933333	0.1697361	0.000000	0.0666667
News	0.820001	0.2711160	0.111500	0.0000000
Advt.	0.550012	0.349901	0.407407	0.157706
Advt.	0.692244	0.300556	0.115223	0.046691
sports	0.221009	0.722140	0.110110	0.000000
sports	0.183302	0.799811	0.001910	0.112001

Table 2: Results: Editing Style(Transition)

magnitude of a tile is computed as follows:

$$\tau_i = \sum_{i=m}^{n} B_i$$

Where τ_i is the i_{th} tile, m is the starting bin of the tile, n is the ending tile and B_i is the value of i^{th} histogram bin. Using the value of tiles as input the FRBS is evolved for characterizing the global shot activity as **high, moderate or low**. A typical high-activity subsampled shot is shown in Fig. 2. It is a crowd scene having lots of activity and therefore categorized as high-activity shot as indicated by membership values.

After computing the activity for each shot we compute the normalized length component for low(nlc_{low}), moderate($nlc_{moderate}$) and high(nlc_{high}) activity for the complete sequence similar to pan and zoom length component computation. These are used as representatives of global activity in the final categorization.

Local Shot Activity Estimation Shot activity can also be categorized in terms of its spatial distribution across the frames. We call this activity as the local activity. We divide each frame into fixed number of regions(in our case its 9). For each region the global activity is computed as mentioned in the previous section. We characterize each region into two categories: high and low, thus getting eighteen variables(for 9 regions) which further go into evolutionary learning for characterizing the local shot activity. The local shot activity is categorized as: localized or distributed. In Fig. 2 activity is distributed in all the regions, therefore it is characterized as distributed activity shot as indicated by the membership values. Normalized length component of localized activity ($nlc_{localized}$) and distributed activity ($nlc_{distributed}$) are then computed, similar to other features.

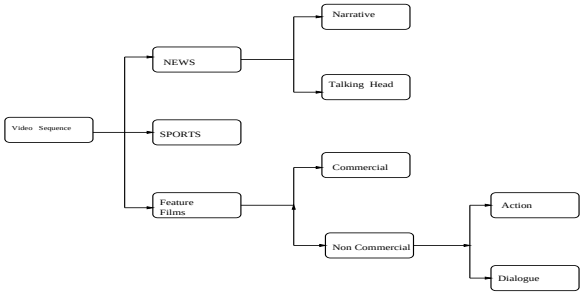

Figure 3: Hierarchical Classification into Generic Categories

4 Fuzzy Rule-based Systems for Video Categorization

In this section we will be presenting an evolutionary learning based approach for classifying the video sequences into generic categories like sports, news and features films. This classification is based on video structure based features extracted in the previous sections. We have evolved fuzzy systems to categorize the video sequences in an hierarchical manner as shown in Fig. 3.

For each level of the hierarchy a separate fuzzy rule based system is being evolved. In the subsequent subsections we will discuss the salient features of these fuzzy systems with illustrations.

4.1 Top level classification: Sports, News and Feature Films

The rules for classification of video sequences into semantic categories are not intuitively obvious. In other words, it is not trivial to identify the combinations of fuzzy features which characterize each category of video. We have explored use of evolutionary learning for identifying appropriate combinations. The choice of training patterns is the key issue for the success of such a learning scheme. We have collected a large number of training samples from various categories like sports, news, documentaries, movies etc. We have tried to evolve the classification systems for

High: 0.850021 Moderate: 0.113211 Low: 0.000290 Localized=0.000001 Distributed: 0.619844

Figure 2: Shot Activity Characterization

various combinations like informative, movie and sports; commercial, non-commercials and sports; movies and others and so on, and finally we could evolve the FRBS which is capable of categorizing a given sequence into three generic categories: **sports, news and feature films**. This categorization seems to be logical because there is a significant difference in terms of feature space among these categories. For example the pace is low in the case of news, moderate in the case of sports and moderate or high in the case of feature films. Similarly camera motion is high in sports while low in news and feature films. Some of the evolved rules for this classification are shown below:

- If $pace$ is slow and $editing - style$ is sharp and nlc_{pan} is small and nlc_{zoom} is small and nlc_{high} is moderate and $nlc_{moderate}$ is small nlc_{low} is large and $nlc_{localized}$ is large and $nlc_{distrubuted}$ is small then it is news sequence.
- If $pace$ is moderate and $editing - style$ is moderately-sharp and nlc_{pan} is large and nlc_{zoom} is very-small and nlc_{high} is large and $nlc_{moderate}$ is moderate nlc_{low} is large and $nlc_{localized}$ is very-large and $nlc_{distrubuted}$ is very-small then it is news sequence.
- If $pace$ is moderate and $editing - style$ is smooth and nlc_{pan} is moderate and nlc_{zoom} is moderate and nlc_{high} is large and $nlc_{moderate}$ is small nlc_{low} is small and $nlc_{localized}$ is moderate and $nlc_{distrubuted}$ is large then it is sports sequence.
- If $pace$ is moderate and $editing - style$ is smooth and nlc_{pan} is very-large and nlc_{zoom} is small and nlc_{high} is large and $nlc_{moderate}$ is moderate nlc_{low} is very-small and $nlc_{localized}$ is small and $nlc_{distrubuted}$ is very-large then it is sports sequence.
- If $pace$ is slow and $editing - style$ is moderately-smooth and nlc_{pan} is large and nlc_{zoom} is large and nlc_{high} is moderate and $nlc_{moderate}$ is large nlc_{low} is moderate and $nlc_{localized}$ is moderate and $nlc_{distrubuted}$ is large then it is sports sequence.
- If $pace$ is fast and $editing - style$ is sharp and nlc_{pan} is moderate and nlc_{zoom} is small and nlc_{high} is large and $nlc_{moderate}$ is moderate nlc_{low} is small and $nlc_{localized}$ is small and $nlc_{distrubuted}$ is very-large then it is feature film.
- If $pace$ is fast and $editing - style$ is sharp and nlc_{pan} is large and nlc_{zoom} is small and nlc_{high} is very-large and $nlc_{moderate}$ is small nlc_{low} is very-small and $nlc_{localized}$ is very-small and $nlc_{distrubuted}$ is large then it is feature film.

Note, that there is a definite correlation among the rules which characterizes a particular category e.g. in news video pace is generally slow, editing style is sharp, camera motion is less and shot activity is low and localized, while in sports pace is moderate, editing style is moderately smooth, camera motion is very high and shot activity is high and distributed. For this classification we have generated a training set consisting of 60 patterns (15 sports, 15 news and 30 feature films). It converged after 3280 generations with a fitness value of 2325.23. Some of the typical results for this classification are shown in Tables 3.

Original	feature	news	sports	classified as
feature	0.999999	0.694959	0.118794	feature
feature	0.999981	0.039362	0.337985	feature
news	0.002703	0.760453	0.000001	news
news	0.000000	0.000000	0.560458	sports
sports	0.000000	0.000000	0.969747	sports
sports	0.173561	0.000000	0.605608	sports

Table 3: Results: Top Level Classification

The overall performance of top level classification is given in Table 4

Seq Type	No. of Sequences In		Correctly Classified In	
Type	Training Set	Test Set	Training Set	Test Set
Sports	20	40	17(85%)	32(80%)
News	15	30	12(85%)	23(76%)
Feature	25	50	21(84%)	39(78%)

Table 4: Performance of Top Level Classification on Training and Test Set

4.2 Classification of Feature Film: Commercial and Non-Commercial

At the highest level advertisements, action movies and other movie sequences are classified in one generic category: Feature Films. We have further made an effort to distinguish between them by evolving another FRBS with the training patterns consisting of only feature films. This is basically a two class problem(commercial and non-commercial). Some of the typical results for this classification are shown in Table 5. In this case we obtain 87% correct classification.

Original	commercial	noncommercial	classified as
noncom	0.484235	0.999965	noncom
noncom	0.129580	0.999990	noncom
com	0.696518	0.562065	com
com	0.941289	0.022658	com

Table 5: Results: Second Level of Classification

4.3 Classification of Non-Commercials: Action and Dialogue

Non-commercial movie sequences are further categorized into: action sequences and dialogue sequences. The learning is done with training patterns taken from non-commercial category alone. Some of the results for this classification are shown in Table 6. The overall correctness obtained for this level of classification is 95%.

Original	dialogue	action	classified as
action	0.781261	0.048634	action
action	0.745809	0.532862	action
dial	0.000653	0.999991	dial
dial	0.000500	0.704317	dial

Table 6: Results: Third Level of Classification

4.4 Classification of News: Talking-Head and Narrative

From a complete news sequence, we extracted that part where news reader is shown or some interview shots are there. We stored this as a separate video sequence and called it as 'talking head video sequence'. Other news clips are kept in the category of 'narrative news sequences'. We created training data set having features from talking head and narrative video sequences alone and learned them. The distinction between these two categories is obvious in terms of feature vector. Talking head sequences have a very high component of localized activity, while camera work and pace is comparatively high in other news sequences. Some of the typical results for this classification are shown in Table 7. The overall correctness obtained for this classification is 87%.

Original	narrative	talking-head	classified as
nar	0.843261	0.430061	nar
nar	0.998658	0.551783	nar
talk	0.918623	0.999996	talk
talk	0.929931	0.966610	talk

Table 7: Results: News Classification

5 Conclusions

In the present work a generic fuzzy theoretic approach for semantic categorization of video sequences using syntactic features has been proposed. Starting with the problem of segmenting the video data, schemes up to content based classification of the video have been suggested. For the purpose of categorization, the use of fuzzy features and fuzzy rules has been explored because of the ability of a fuzzy framework to accommodate the variations inherent in the different categories of the video sequences.

References

[1] D. Bordwell and K. Thompson. *Film Art, 5th Edition*. McGraw-Hill, 1997.

[2] C. Dorai and S. Venkatesh. Bridging the Semantic Gap in Content Management Systems: Computational Media Aesthetics. In *Proc. Conf. on Computational Semiotics for Games and New Media, 10-12 Sept., CWI Amsterdam*, pages 94–99, 2001.

[3] N. Haering, R. J. Qian, and M. Ibrahim. A Semantic Event-Detection Approach and Its Application to Detecting Hunts in Wildlife Video. *IEEE Transactions on Circuits and Systems for Video technology*, 10:857–868, September 2000.

[4] N. Hirzalla and A. Karmouch. Detecting Cuts by Understanding Camera Operation for Video Indexing. *Journal of Visual Language and Computing*, 6:385–404, 1995.

[5] B. Horn and B. Shunk. *Robot Vision*. MIT PRESS, 1986.

[6] I. Ide, R. Hamada, H. Tanaka, and S. Sakai. News Video Classification based on Semantic Attributes of Caption. In *Proc. 6th ACM International Conference*, pages 60–61, 1998.

[7] R. Jadon, S. Chaudhury, and K. Biswas. A Fuzzy Theoretic Approach for Video Segmentation Using Syntactic Features. *Pattern Recognition Letters*, 22:1359–1369, November 2001.

[8] R. Jadon, S. Chaudhury, and K. Biswas. A Fuzzy Theoretic Approach for Camera Motion Detection. In *Proc. IPMU-2002, Annecy, France, July 1-5*, 2002.

[9] V. Kobla, D. Doermann, and C. Faloutsos. Developing High-Level Representations of Video Clips using Video Trails. In *Proc. SPIE Conference on Storage and Retrieval of Image and Video Databases VI*, pages 81–92, January 1998.

[10] M. Roach, J. Mason, and M. Pawlewski. Video Genre Classification Using Dynamics. In *Proceedings ICASSP-2001, May 7-11, Salt Lake City, Utah*, 2001.

[11] Y. Shi, R. Eberhart, and Y. Chen. Implementation of Evolutionary Fuzzy Systems. *IEEE Transactions on Fuzzy Systems*, 7(2):109–118, April 1999.

[12] N. Vasconcelos and A. Lippman. Statistical Models of Video Structure for Content Analysis and Characterization. *IEEE Transactions on Image Processing*, 9(1), January 2000.