

# Fuzzy Multimedia Mining Applied to Video News

**Marcin Detyniecki**

CNRS, Pôle IA, LIP6,  
Université P. et M. Curie - Paris 6  
8 rue du Capitaine Scott,  
75015 Paris, France  
*Marcin.Detyniecki@lip6.fr*

**Christophe Marsala**

Pôle IA, LIP6  
Université P. et M. Curie - Paris 6  
8 rue du Capitaine Scott,  
75015 Paris, France  
*Christophe.Marsala@lip6.fr*

## Abstract

In this paper, we present the application of the general scheme of Knowledge Discovery to Multimedia data. In particular, we discuss the extraction of structures from video news. Concrete results are presented for the extraction of knowledge based on color properties (of characteristic video images) using a fuzzy decision tree based method.

**Keywords:** Multimedia mining, Video, Fuzzy decision trees.

## 1 Introduction

On the one hand, the growth of video data has caused a corresponding need to analyze and exploit them. Perfect examples of this increase are the availability of video news on the web, or the selling of video tape recorders that saves video data directly on a hard drive. Traditionally, it appears that the user tends to interact with this kind of objects in order to obtain what he wants or even to personalize it. Unfortunately, today, this can be only done by using elementary text queries. Furthermore, the general approach is an information retrieval context, where the aim is to find specific and previously indexed informations and not a new kind of knowledge [14].

On the other hand, in the recent years, fuzzy data mining introduces new methodologies to extract and discover fuzzy knowledge from either classical or fuzzy data repositories. It

leads to the improvement of the knowledge discovery process that enables us to offer more comprehensive discovered knowledge, and to enhance its capabilities to handle numerical or fuzzy data [16]. Thus, it appears natural and promising to link fuzzy data mining with multimedia data that leads to fuzzy multimedia mining, an extension of the recent multimedia data mining [17]. For instance, an application resides in mining video news to extract meaningful and useful information (topic of a sequence, persons involved in a video,...), and providing the user with either personalized or browsable (structured) news. Here, a new problematic arises since a lot of informations can be extracted from video: texts, images, sounds, temporal data, metadata,...

A solution lies in providing a flexible and automated data-mining tool which will induce knowledge from all kinds of data. A particular instance of such tools is the fuzzy decision tree based algorithm [16].

It can be noted that video data indexing is a kind of extraction process already known in multimedia. Existing literature on video indexing implicitly defines video indexing as the process of extracting the temporal location of a feature and its value from video data [6]. However, indexing is generally done manually, and the growth of video data and the demand for new applications with finer grain access to video, highlight the fact, that an automation of the indexing process becomes essential.

In this paper, we first present in Section 2 the general framework of knowledge discovery in the case of multimedia systems. We

present the adapted scheme going from multimedia data to knowledge. We also present tools available to fulfill this goal. In Section 3, we focus on the particular case where the knowledge to be extracted is the structure of a video news. Section 4 illustrates practically the versatility of our approach by mining color features from key-frames.

## 2 Multimedia data mining

Knowledge Discovery from Data (KDD) was introduced by Fayyad, Piatetsky-Shapiro and Smyth at the beginning of the nineties [9]. Taking into account the polymorphism of multimedia data, Multimedia Data Mining (MDM) was recently proposed as a new topic of research [13, 17, 26, 22].

Since mining is completely linked to the kind of data, a basic approach is to separate the different media channels (Figure 1). By do-

ditional steps: the extraction of the various media (spatial, temporal, audio...) [2] at the beginning of the process in order to highlight various streams of data, and eventually the aggregation of the specialized knowledge obtained from each stream.

**Visual spatial content.** A representative image can be easily extracted from the video stream. From this image, several features can be extracted.

The *color* feature of an image is typically represented by histograms. The *texture* feature describes the contrast, the uniformity, the coarseness, the roughness, the frequency, the directionality, in the image. The *sketch* feature gives a representation of the image containing only outlines of objects. The *shape* feature describes global features in the image as the circularity, eccentricity and major axis orientation, but also local ones such as for instance point of curvature, corner location, turning angles and algebraic moments.

**Visual temporal content.** Another important characteristic of a video is its temporal aspect.

The *camera* motion describes the real and factual movement of the camera. It is usually obtained either by studying the optical flow by dividing the video in several regions [24] or by studying the motion vector [18]. The *object* motion describes the trajectory obtained by tracking one object on the screen [15, 21].

**Audio content.** From the audio stream, we can extract the following basic characteristics.

The *loudness* is the strength of the sound, determined by the amplitude of the sound wave. The *frequency* translates what we perceive as pitch. The *timbre* is the characteristic distinguishing sounds from different sources.

More sophisticated characteristics can be obtained as for instance speaker/tracking and noise, silence and speech segmentation.

Another interesting feature is to recognize when *noise*, *music*, *speech*, or *silence* is predominant. Various approaches exist such as

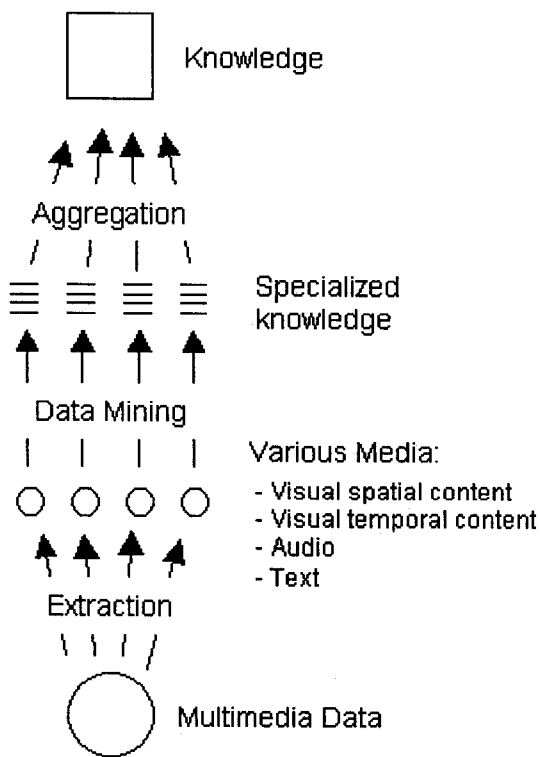


Figure 1: Multimedia data mining: an approach

ing this, the classical knowledge discovery scheme is modified by introducing two ad-

the use of expert systems [5] or hidden Markov chains [1].

**Text content.** The text can bring out very useful knowledge. A lot of research has been done in order to extract (or synchronize) some text when it is not available. Another interesting challenge is to extract the written information appearing on the screen. The idea is to locate the text on the screen and then to recognize it. It is to notice that very often the synchronized text is available and can be exploited (for instance subtitles on a special channel).

### 3 Discovering structures in video news

In order to extract structures from video news, we instantiate the multimedia knowledge discovery scheme (Figure 2). Here the

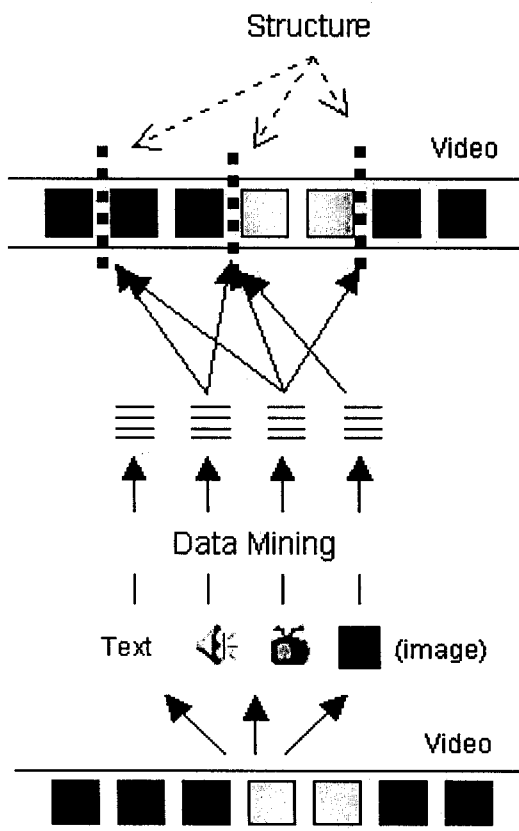


Figure 2: Finding structures in video

multimedia input is the video containing the

news. The extracted contents are the text, the audio, the camera movements and representative frames (images).

By analogy with existing methods in distributed data mining [12, 23], we proposed here an approach to mine each of these media separately. The main idea is to extract knowledge from the whole kind of available data and to bring up a collaborative process where each kind of data will help to extract a global structure (see Figure 2).

For instance, in video news, changes of subject (and sometimes the subject itself) of reports can be extracted from the spoken text. Using the audio file, significant silence (music or/and voice) periods should be highlighted, that may indicate wanted pauses of the news host (or editor) underlining the news structure. The type of camera motion may also be useful to confirm or recognize the type of shot<sup>1</sup>: host talking, violent movement of crowds, war scenes,...

Finally, representative images of each shot (ie. *key-frames*) can be used to discover if the host, a map or a correspondent appears on the screen. Such information enables us to understand the structure of the news. We can also detect the key-frames containing inlayed texts that have some important information (that could be extracted) and that announce the beginning or the end of a semantic unit.

**Fuzzy decision trees.** In a multimedia framework, a versatile data-mining tool is necessary. In our application, we use a fuzzy decision tree learning algorithm in order to obtain rules that summarize and explain the data: the software Salammbô is able to handle numerical input, and constructs fuzzy decision trees [3, 16]. The advantage of using fuzzy decision trees resides in the fact that they represent a natural and understandable knowledge: a fuzzy decision tree is equivalent to a set of fuzzy IF...THEN rules.

<sup>1</sup>We recall that a shot is a sequence of images on which there is no change of camera.

## 4 Mining color features

To illustrate one of the branches of the multimedia mining scheme (Figure 2), we focus here on the mining of knowledge from video-news key-frames.

### 4.1 Extraction of key-frames

In this application, we start from a set of key-frames extracted per shot. Each image of the set contains characteristic features (colors) from the sequence of video frames in one shot. The advantage of working with key-frames is to reduce the image processing to one image per shot (against 25 times length in seconds of the shot).

Classical techniques used to find shots are in the uncompressed domain based on pixel-wise or histogram comparison [10]. In the compressed domain [27] they are based either on coefficient manipulations as inner product or absolute difference or on the motion vectors. Then the key-frames are usually extracted with the simple heuristic that consists in taking the first or the tenth frame [27]. More sophisticated methods look for local minima of motion or significant pauses [25].

In our application, we use previous works realized by the multimedia group of the LIP6 to work on the uncompressed domain [7, 8, 19, 20].

### 4.2 Vectorization of key-frame colors

In a second step, the set of colors from a key-frame is vectorized in order to obtain a global set of vectors to compare key-frames.

Given a reference palette of colors (for instance a palette of 64 or 256 colors obtained by discretizing equally the RGB space), an histogram of the frequencies of each color is computed for each key-frame (here, a key-frame is an image in gif format with its 256 specific colors). It enables us to obtain a vector in a reference space (defined by the colors of the reference palette) for each key-frame. To compute this vector, each original color in the palette of a key-frame is projected (using

a euclidean distance) in the HSV color-space to the reference palette of colors (Figure 3).

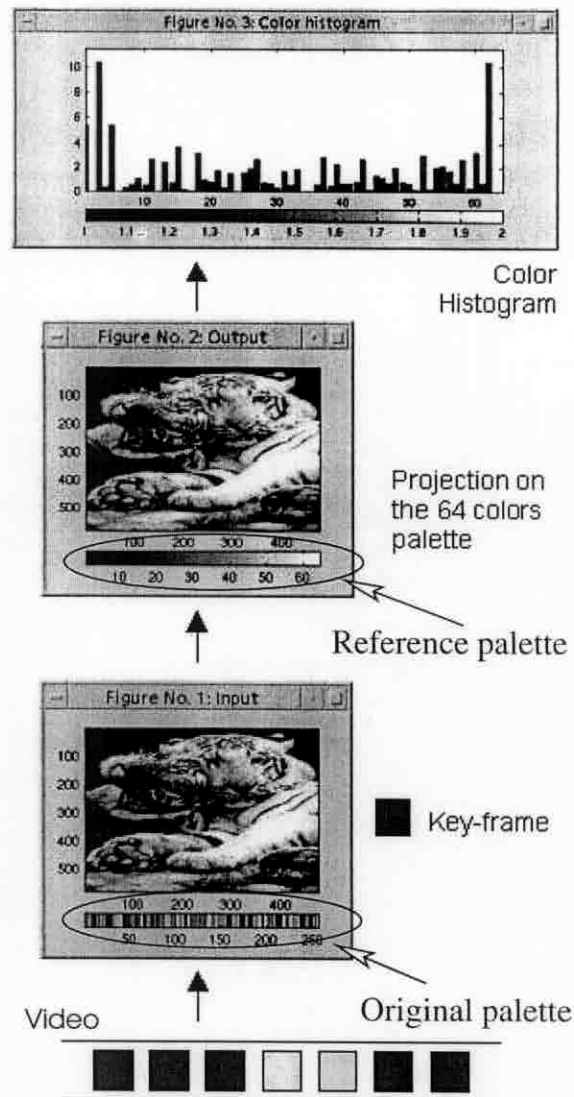


Figure 3: Color palette transformation

### 4.3 Experimentations

Each key-frame enables us to construct a vector which can be considered as a training example. Thus, from an original set of key-frames, a training set can be composed to feed the Salammbô software and give rise to a fuzzy decision tree.

In order to show the versatility of the approach, three different mining problems have been considered (all related to the extraction of the structure). In the following we present

these problems and the obtained results.

### 4.3.1 Discovering the presence of inlays in a key-frame

Inlays that appear on the TV screen are very often hints for the structure of the video news. They often appear when a new person is presented or when a report related to a subject will end (names of the correspondents). They usually consist in a square or a rectangle that frames some text. We assume that one color (or color proportions) usually remains the same over the journals of one channel.

We have conducted several mining experiments in order to determine if colors are discriminant for the detection of inlays. A training set has been composed with 176 vectorized key-frames, each vector has been assigned to the class (with or without inlays) of the key-frame.

A first experiment was conducted with the whole training set and the reference palette composed by 64 colors. The root of the fuzzy decision tree constructed is the white color attribute (the major background color of inlay key-frames). The first part of this tree is given in Figure 4. This result points out that only a

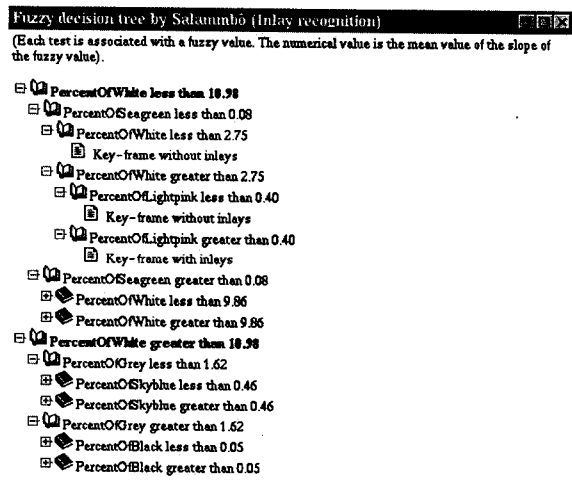


Figure 4: Part of the fuzzy decision tree for the recognition of inlays

few number of colors is needed to discriminate inlay key-frames.

Taking into account this result, a second ex-

periment was conducted to evaluate the importance of the number of colors of the reference palette. A training set was constructed with the 176 vectorized key-frames for each size of reference palette (16, 64, 256 or 512 colors). A cross-validation was conducted by splitting the training set into 4 subsets, each one with 22 examples of each class. A training was done using simultaneously 3 of these 4 sets, the fourth one was used as a test set, and so on. Results are presented in Table 1.

Table 1: Inlay recognition with various palette sizes.

Palette size	Accuracy
16 colors	64.2%
64 colors	81.25%
256 colors	79.55%
512 colors	79.55%

They highlight the fact that only a minimal set of colors is required. It can be explained by the fact that an inlay is often a big pattern with a predominant color, thus there is no need in increasing the colors of the palette to obtain better results in accuracy. However, there is a minimum size for this palette: a few number of colors will lead to wrong projection to the reference palette.

Finally, a third experiment was conducted to position the Salammbô algorithm with regard to other learning algorithms<sup>2</sup> for the detection of inlays. In Table 2, the accuracy to recognize inlay key-frames or normal key-frames of each tested algorithm is indicated. We present also their precision and recall values for each kind of key-frames, and the building time<sup>3</sup> of the learning model. In this kind of applications, the recall and precision values of the model are often as important as the accuracy of the model: it is important to perfectly recognize at least one kind of key-frames. We can observe that the fuzzy de-

<sup>2</sup>These tests have been done thanks to the free software Weka from the University of Waikato (NZ) (<http://www.cs.waikato.ac.nz/ml/weka/>).

<sup>3</sup>Building times are given in seconds and are valued with a Sun Ultra 5 station.

Table 2: Results for inlays recognition (64-colors palette).

Method	Accuracy	With inlays		Without inlays		Building time
		Recall	Precision	Recall	Precision	
Salammbô	81.25%	0.875	0.778	0.75	0.857	1
Naive Bayes	54.55%	0.386	0.567	0.705	0.534	0.4
Voted Perceptron	71.59%	0.705	0.721	0.727	0.711	0.6
Weka J48 (C4.5)	78.41%	0.75	0.805	0.818	0.766	1
Decision Table	82.95%	0.864	0.809	0.795	0.854	3.3
AdaBoost (Weka J48)	85.8%	0.864	0.854	0.852	0.862	5.6
Neural Network	80.11%	0.841	0.779	0.761	0.827	322

cision tree based method is not only among the methods with the higher accuracy, but also, presents high recall and precision rates for inlays recognition. Moreover, it appears that the construction of fuzzy decision trees by Salammbô is also among the low time consumer methods, an important property for multimedia applications. This, in addition of the understandability of the fuzzy decision tree model, Salammbô presents better accuracy and quickness ratio than other presented numerical methods.

#### 4.3.2 Discovering errors in the shot detection

The shot detection algorithms are usually very sensitive. The idea is to be sure not to miss any shot change. The drawback of this approach is the over-segmentation of the video. For instance, the appearance of inlays is often the cause of splitting a single shot into two separate shots.

Thus, we have conducted an experiment to discover when two successive key-frames are part of the same shot. Here, the training set was constructed as follows. A first group of training examples was composed by two successive key-frames from the same shot. A second group was composed by two successive key-frames from different shots. Each key-frame is vectorized in a 64 colors palette, the two vectors are merged to obtain a single training example with 128 features.

At this point, we expect to find rules translating the similarity. But, the built fuzzy de-

cision tree highlights that the most discriminant approach is to detect the increase of the proportion of a particular color: the dominant color of inlays. This rule suggests that after a classical shot detection (based on similarities), the accuracy of the shot detection can be enhanced using the color dissimilarity between successive key-frames.

In this experiment, the mean accuracy with cross-validation of fuzzy decision trees is 78.26%.

#### 4.3.3 Discovering host, diagram, correspondent

An important hint about the structure of the news is to detect the appearance of either the host, an explanatory diagram, or a correspondent (not in the studio). Thus a last experiment was conducted in order to recognize the presence or non-presence of the host speaker in a key-frame.

As in previous experiments, a training set was constructed. A first group of training examples was composed by key-frames where the host appears. A second group was composed by key-frames without host. Each key-frame is vectorized in a 256 colors palette which is the best size for this experiment.

The most effective fuzzy rule to recognize if we are in presence of the host (or not) is to detect the color of the host's background (presently, one blue color). This rule points out that the best way to know if we are in presence of the host is to look if the scene takes place in the channel studio. The accuracy of the fuzzy de-

cision tree here is 88%.

#### 4.4 Discussion

For this three different problems, the extracted knowledge is in the form of three seminal rules. For the first one, the system suggests to detect a large proportion of a single color, putting forward that all inlays have similar colors.

This basic problem and its result have been used to conduct some experiments relative to colors and some comparisons with other learning systems have been done, which point out that the Salammbô software, is particularly successful taking into account time, accuracy, recall and precision.

For the other two problems our knowledge discovery system extracted two astounding rules:

- in order to detect the presence of a host, focus on the background, which corresponds to discern if the scene was taken in the studios. So, this rule suggests not to look at the host to detect the presence of a host.
- in order to ameliorate the shot detection (which is naturally based on similarity) it is recommended to look at the differences between key-frames. More precisely, the rule suggests that an inlay has been detected, implying at the same time that this is actually the cause of the errors.

#### 5 Conclusion

In this paper, we presented the general scheme of the multimedia knowledge discovery in the case of multimedia systems. We applied it to the extraction of knowledge in the form of video-news structures. We used fuzzy decision trees, because of the simplicity and the understandability of the extracted rules.

Practically we focused on key-frame color mining in order to notice important structural hints as for instance the appearance of important informations on the screen, the presence of the host, or also correcting possible mistakes in the shot detection.

This is a first step in the multimedia mining of a video. The next step is to continue the mining of visual contents as for instance the texture and also the mining of structural content. We will further mine the other medias (sound, text, etc.) in order to complement the hints in a final step.

Other future work will consider a more global approach to mine the media jointly. Exploitation and combination of information from each tool will be done during the process of data mining and not only at the end of each process.

#### Acknowledgements

The authors wish to thank here M. Gwenaël Durand for providing an indexed record of video news.

#### References

- [1] R. André-Obrecht. Special Issue on Speaker Recognition and its Commercial and Forensic Applications. In *Int. Jour. Speech Communication*, North Holland, Vol. 31, Nr. 2-3, June 2000.
- [2] Y.A. Aslandogan and C.T. Yu. Techniques and Systems for Image and Video retrieval. In *IEEE Transactions on Knowledge and Data Engineering*. Vol. 11, No. 1, 1999, pp. 56-63.
- [3] B. Bouchon-Meunier, C. Marsala and M. Ramdani (1997). Learning from Imperfect Data. In *Fuzzy Information Engineering: a Guided Tour of Applications*, D. Dubois, H. Prade and R. R. Yager eds, chapter 8, pp. 139-148, 1997.
- [4] M. Davis (1993). Media streams: An iconic visual language for video annotation. In *IEEE Symposium on Visual Languages*, pp. 196-202. IEEE Computer Society, 1993.
- [5] M. De Santo, G. Percannella, C. Sansone and M. Vento (2001). Classifying Audio Streams of Movies by a Multi-Expert System. In *Proc. of Int. Conf. on Image Analysis and Processing (ICIAP01)*, Palermo, Italy, Sept. 26-28, 2001.
- [6] M. Detyniecki and C. Marsala (2001). Fuzzy inductive learning for multimedia mining. in *Proc. of the EUSFLAT'01 Int. Conf.*, pp. 390-393, Leicester, UK, Sept. 2001.

- [7] G. Durand and P. Faudemay. Cross-indexing and access to mixed-media contents. in *Proc. CBMI'01 - International Workshop on Content-Based Multimedia Indexing*, Brescia, Italy, Sept. 2001.
- [8] G. Durand, C. Thienot and P. Faudemay. Extraction of Composite Visual Objects from Audiovisual Materials. in *Proc. SPIE - Multimedia Storage and Archiving Systems IV*, Vol 3846, pp. 194-203, Boston, Sept. 1999.
- [9] U. M. Fayyad, G. Piatetsky-Shapiro and P. Smyth (1996). From Data Mining to Knowledge Discovery in Databases. In *AI Magazine*, 17:3, pp. 37-54, 1996.
- [10] F. Idris and S. Panchanathan. Review of image and video indexing techniques (1997). In *Journal of Visual Communication and Image Representation*, 8:146-166, 1997.
- [11] P. Joly (1996). Consultation et analyse des documents en image animée numérique. *Thèse de l'Université P. Sabatier - Toulouse*, 1996.
- [12] H. Kargupta, B.-H. Park, D. Hershberger, E. Johnson (1999). Collective Data Mining: a New Perspective Toward Distributed Data Analysis. in *Advances in Distributed and Parallel Knowledge Discovery*, H. Kargupta and P. Chan eds. MIT/AAAI Press. 1999.
- [13] R. Kosala and H. Blockeel (2000). Web Mining Research: A Survey. In *SIGKDD Explorations*, volume 2, issue 1, pp. 1-15, 2000.
- [14] D. Kraft, G. Bordogna and G. Pasi (1999). Fuzzy Information Retrieval, in *International Handbook of Fuzzy Sets*, J.C. Bezdek, D. Dubois and H. Prade eds., Kluwer Academic Pub., Vol.3, chapter 6, pp. 469-510, 1999.
- [15] S.Y. Lee and H.M. Kao (1993). Video indexing - an approach based on moving object and track. In *SPIE*. 1908:81-92. 1993.
- [16] C. Marsala and B. Bouchon-Meunier (1999). An Adaptable System to Construct Fuzzy Decision Trees. In *Proc. of the NAFIPS'99*, New-York, pp. 223-227, June 1999.
- [17] First Workshop of Multimedia Data Mining (2000). <http://www.cs.ualberta.ca/zaiane/mdm.kdd2000/>, 2000.
- [18] M. Pilu (1997). On using raw MPEG motion vectors to determine global camera motion. *Tech. report of Digital Media Dept. of HP Lab. Bristol*, Aug. 1997.
- [19] R. Ruiloba and P. Joly (2000). Framework for evaluation of video-to-shots segmentation algorithms. In *Video Data special issue of Networking and Information Systems*. Vol. 3, pp. 46-57. 2000.
- [20] R. Ruiloba, P. Joly, S. Marchand-Maillet and G. Quénot (1999). Towards a Standard Protocol for the Evaluation of Video-to-Shots Segmentation Algorithms. In *Proc. of the European Workshop on Content Based Multimedia Indexing*, pp 41-48, Toulouse, France, 1999.
- [21] E. Sahouria (1997). Video indexing based on object motion. *M. S. thesis*. U.C. Berkeley. 1997.
- [22] S.J. Simoff (2000). Variations on Multimedia Data Mining. In *Proc. of the First International Workshop on Multimedia Data Mining (MDM/KDD'2000)*, O.R. Zaïane and S.J. Simoff eds, Boston, USA , pp. 104-109, August 2000.
- [23] S. J. Stolfo, A. L. Prodromidis, S. Tselepis, W. Lee, D. W. Fan, and P. K. Chan. Jam: Java agents for meta-learning over distributed databases. In *Proc. of the 3rd Int. Conf. on Knowledge Discovery and Data Mining*. Newport Beach, CA, pp. 74-81 August 1997.
- [24] G. Sudhir and J.C.M. Lee (1996). Video annotation by motion interpretation using optical flow streams, *Journal of Visual Communication and Image Representation*. 7:354-368. 1996.
- [25] H.H. Yu and W. Wolf (1999). A hierarchical multiresolution video shot transition detection scheme. *Computer Vision and Image Understanding*. 75:196-213. 1999.
- [26] O.R. Zaïane, J. Han, Z.-N. Li, S.H. Chee and J.Y. Chiang (1998). MultiMediaMiner: A System Prototype for Multimedia Data Mining. In *Proc. of the ACM Sigmod, Int. Conf. on Management of Data*, pp. 581-583, 1998.
- [27] H.J. Zhang, C.Y. Low, S.W. Smoliar and J.H. Wu (1995). Video parsing, retrieval and browsing: an integrated and content-based solution. In *Proc. of ACM Multimedia 95 - Electronic Proceedings*. San Francisco, CA, Nov. 1995.