CARTOON DETECTION USING FUZZY INTEGRAL[1]

Antonio Rama, Francesc Tarrés, Laura Sanchez

Dept. Teoria del Senyal i Comunicacions - Universitat Politècnica de Catalunya, Barcelona, Spain

ABSTRACT

With the growth of digital television TV program classification has become a major research topic. Recent classification techniques have reported acceptable results for specific genre detection. Cartoons is one of these genres which has deceived some attention because of its importance in push scenarios where parents want to control their children's access to television. In this paper a flexible scheme based on a non-linear classifier called Fuzzy Integral is presented. This operator is supposed not only to classify but also to give a relevance measure to all the features involved in the classification. Preliminary results using this operator for cartoon detection are presented and compared with other well known statistical clasification methods like PCA, LDA or K-NN.

1. INTRODUCTION

TV genre classification is a fundamental component in many applications for this new multimedia world. Just to give a couple of examples let us consider a video indexing system for generating automatically metadata that enables fast browsing and retrieval, or some modules which can filter or select predetermined TV genres for digital television set-top-boxes. This push scenario is very useful for locking violent and sexual content to children, for automatic channel switch to predefined TV programs or for recording commercial-free video material.

In the current literature, the approaches used in TV content analysis have evolved from single modality techniques, where either audio or video features are used, to multimodal algorithms where both audio and video features are combined to make decisions. In multimodal analysis the problem of selecting the optimal subset of features is more important due to the huge number of available measures. Thus, it is very important to select only those features that give us the most relevant information for the classification problem. Therefore, only those measures presenting the most significant information will be selected for the classification This paper presents a fusion operator of video and audio data called Fuzzy integral that together with classifying the shots in the different TV programs (sports, news, commercials, music clips, others) it also quantifies the relative importance of each feature, separately and combined with others, in the final classification. Thus, it will be possible to detect the audiovisual features that are most discriminative for a determined TV genre allowing to implement fast TV shot selection modules for a given TV genre. In this paper only the discriminative power of the Fuzzy Integral operator is considered, showing that higher performance over other conventional state-of-the art classfications methods is obtained .

The rest of the paper is organized as follows. In section 2 an overview of some recent work in multimodal content analysis is briefly presented with special emphasis in cartoon detection. Section 3 summarizes the fundamentals of the Fuzzy Integral operator whereas section 4 reports about the material used for testing the Fuzzy Integral operator and the experiments performed. And finally some results, conclusions and future work are exposed in sections 5 and 6 respectively.

2. RECENT WORK IN CARTOON DETECTION

In the last years, the researchers have realized the benefits of using audio and video features jointly for the analysis of video content. These recent multimodal approaches have reported better performance than using audio or video analysis separately. [Fischer95] was one of the pioneers that investigated automatic recognition of film genres based on multimodal analysis. The authors classified the TV programs into news cast, car races, tennis, commercials or animated cartoons by using three levels of abstraction. [Truong00] presented a set of computational features originating from a study of editing effects, motion, and color used in videos, for the task of automatic video categorization. Classification results from experiments on several hours of video established recognition rates from 80% to 83% when differentiating

[1] ACKNOWLEDGMENT: The work presented was developed within VISNET-II, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 programme

between cartoons, commercials, music, news and sports. Xu et al. [Xu03] investigated the problem of automated video classification by analyzing the low-level audio-visual signal patterns along time in a holistic manner. Five popular TV broadcast genre were studied: sports, cartoon, news, commercial and music. A novel statistically based approach was proposed comprising two important ingredients designed for implicit semantic content characterization and class identities modeling. First, a spatial-temporal audio-visual "concatenated" feature vector is composed, aiming to capture crucial clip-level video structure information inherent in a video genre. Second, the dimensionality of the feature vector was reduced by eliminating spatial-temporal redundancy using Principal Component Analysis. The result of this stage is a compact probabilistic model for each class genre. An average correct classification rate of 86.5% was achieved, being the sport genre the one with the highest accuracy (96.7%) and cartoons the one with the lowest (79.5%).

More recently, Glasberg et al. [Glasberg05] have presented a new approach for classifying MPEG-2 video sequences as 'cartoon' or 'non-cartoon' by analyzing specific color, texture, and motion features of consecutive frames in real-time. In the proposed method, the extracted features from the visual descriptors are non-linearly weighted with a sigmoid-function and combined using a multilayered perceptron to produce a reliable recognition. The authors reported an average correct classification rate of 80% for cartoon-videos detected as a 'cartoon' and more than 85% for the other genres.

Usually, for all the approaches, cartoon has a detection rate above 80%. Thus, the main objective of this paper is to present a promising tool that achieves this rate using only 8 audiovisual features at audio frame level (each 21 ms), and that could also measure the relevance of the features in the classification process. Next section presents the fundamentals of the fuzzy-integral operator.

3. OVERVIEW OF THE FUZZY INTEGRAL

3.1. Fuzzy Integral fundamentals

The theory of *Fuzzy Measures* is based on the work of Sugeno [Sugeno74]. The introduction of fuzzy sets [Zadeth65], which result from a generalization of classical sets, encouraged the redefinition of set measures. Sugeno achieved this definition by introducing the so-called fuzzy measures, with respect to which fuzzy Integral can be defined. Thus, fuzzy measures generalize classical measures, i.e. probability measures. Here only a brief overview of how fuzzy integral can be used for classification problems is presented and the reader is addressed to [Soria04] for more precise details. The main idea is to use a *fuzzy integral* as a classifier using an extended set of audio and video features in TV program genre classification. Fuzzy Integrals are generalizations of integral operators that include non-linear operations on the data set. In the context of classification, the most frequently used fuzzy integrals are the *Choquet* integral and the *Sugeno* integral. In this work we propose to use a *Choquet* integral for the data fusion process. The main ideas and the process of computing the *Choquet* integral are given hereafter:

Consider we have a vector of feature attributes $X = \{x_1, x_2, \ldots, x_n\}$ where x_i may represent a pixel, an audio sample or (as in our case) an audio or video feature. Given this set of features we collect a number of M samples for the training stage. The attributes of the features at each sample are represented by a vector

$$f = \{f(x_1), f(x_2), \ldots, f(x_n)\} \quad \textbf{Eq 1}$$

The *Choquet* integral is defined in terms of a finite number of fuzzy measures which represent the *a priori* importance of the feature attributes. This set of fuzzy measures is defined for any individual feature and all their possible combinations. Namely, the fuzzy measures are:

$$\mu(\{x_1\}), \mu(\{x_2\}), \mu(\{x_3\}), \ldots\ldots, \mu(\{x_1, x_3\}), \ldots., \mu(\{x_1, x_2, x_3\}), \ldots\ldots \quad \textbf{Eq 2}$$

The *Choquet* integral consist in a two stage process:

1) Rearrangement of the feature values vector in non decreasing order, such that

$$f(x_1') \le f(x_2') \le \cdots \le f(x_n') \quad \textbf{Eq 3}$$

where $(x_1', x_2', \ldots, x_n')$ is a certain permutation of $(x_1, x_2, \ldots, x_n)$.

2) The Choquet integral is then obtained by computing:

$$\int f \cdot d\mu = \sum_{i=1}^{n} \left[f(x_i') - f(x_{i-1}') \right] \cdot \mu\left(\{x_i', x_{i+1}', \ldots x_n'\} \right) \quad \textbf{Eq 4}$$

The training of the classifier consists in selecting the optimum fuzzy measures on the objective of minimizing the misclassification rate. There are a number of alternatives for estimating the fuzzy measures but most of them are based on soft-computing strategies. One of the interesting peculiarities of the *Fuzzy Integral* as a classifier is that once the fuzzy measures have been determined, the classification is computationally very efficient. Moreover, from the values of the fuzzy measures, the *importance* of each feature in the classifier may be found. This result can help to reduce the computational burden required to compute the feature set.

3.2. Implementation of the Fuzzy Integral for cartoon detection

In this work, we are following an approach based on *neural networks* for estimating the set of fuzzy measures. Since we are using different audio visual measures with completely different dynamic ranges, the first step is to normalize the data due to the power of each measure:

$$f(x_i) = x_i \cdot w_i \qquad where \qquad w_i = \frac{1}{M}\sqrt{\sum_{j=1}^{M} x_i^{\,2}} \qquad \textbf{Eq 5}$$

The optimum fuzzy measures are computed using a learning algorithm based on the following control equation:

$$\mu^{i+1}(\overline{f(x')}) = \mu^{i}(\overline{f(x')}) + \sigma \cdot error \cdot \overline{\Delta f(x')} \qquad \textbf{Eq 6}$$

where $\overline{f(x')}$ are the feature values rearranged in non-decreasing order as mentioned in Eq. 3, σ is the adaptive step size and *error* is a parameter that can take the values -1, 0 or 1 depending on the decision of the classifier (0 means that the sample has been correctly classified). And finally $\overline{\Delta f(x')}$ is the difference between all the attributes involved in the fuzzy measure we are updating.

In the next section only the discrimination performance of the operator will be tested assuming that an optimal subset of audiovisual features is given. The results are compared with other well known classification methods of the literature.

4. CARTOON DETECTION RESULTS

4.1. A/V Material description

The available audiovisual material used for the experiments described in the next section consists of 2 hours of A/V material from the "*Televisió de Catalunya*" originally recorded in MPEG2 SD at High Quality (720x576 @ 10Mbps). These 2 hours of video sequences corresponds to almost one hour of cartoons and one hour of other genres like sports, news, commercials, soap opera and documental. This material has been divided into four different sets:

Set A: Non cartoon sequences with a total length of all the clips of 45 minutes approximately.

Set B: Non cartoon sequences with a total lenght of 45 minutes.

Set C: Video sequences of different cartoon types. The total length of all the clips is 10 minutes approximately.

Set D: Same as set C but with different cartoon sequences and a total lenght of 15 minutes.

4.2. Classifiers and Experimental Description.

Apart from the Fuzzy Integral operator described in section 3, three well known classifiers will be used to compare the results. The other three classifiers are a Mean Nearest Neighbour (M-NN), Principal Component Analysis (PCA) and Linear Discriminant Analysis classifier (LDA) [Duda01]. These four classifiers will be used to differentiate between cartoon and non-cartoon at the audio frame level, i.e. each 21.33 ms approximately (1024 samples * 48000^{-1} s/sample). Note that this classification is very hard since the temporal window is less than one video frame and no past or future information is used. Therefore, the methods presented here should be considered only as a first classification step of the data. We will have decisions at audio frame level that are expected to be noisy. Further filtering will be needed, as a second classification step, to decide the clip genre from the output of the first level classifiers. Several approaches can be taken: averaging, median filtering, majority voting, etc. However, as our purpose in to show the discriminant power of the fuzzy integral we only will consider the performance of the first stage of the classifier module.

Another consideration is that we have made a pre-selection of features in order to reduce the number of measures involved in the classification. It is very important to work with a reduced number of features since many of the applications of video genre classification have to be made in real time where reduced computational cost is a must. The main purpose of this approach is to determine reduced sets of features which can be effectively used for TV genre classification. Thus, we decided to use a maximum of 8 features, which have been selected from a total set of more than 30 audio and video measures. The criterion for selecting these features has been an empirical study of the statistics of the feature vector. When using several features at a time, the feature vector autocorrelation matrix *C* provides some measure of the redundancy between the features:

$$C = \frac{1}{N}\sum_{x\in\chi}(x-m)(x-m)^T \qquad with \qquad m = \frac{1}{N}\sum_{x\in\chi} x \qquad \textbf{Eq 7}$$

where χ is the set containing all feature vectors derived from training sequences, and *N* is the total number of feature vectors in χ.

Experiments showed that two highly correlated features will produce similar results than using only one of the features. Moreover, the combination of two high discriminative features that are highly correlated reports worse final results than the combination of a high and low discriminative feature that are slightly correlated. However, the analytic relationship between discrimination power and correlation is difficult to obtain for a general classifier. Moreover, other factors like computational cost should be taken also into account when selecting the feature vector. We have used this empirical criterion for selecting the following feature set: 2 Audio Frame Level Features (Pitch, Roll-off), 3 Audio Clip Level Features (Frequency-Centroid Mean, BW-Mean, Energy SubBand Ratio 1-Mean) and 3 Video Frame Level Features (Entropy HSV, Autocorrelogram, Kurtosis of the Phase Correlation Function).

4.3. Results.

As stated in the previous section, all the results presented here used only 8 features for the multimodal classification which is made at the frame level (each 21.33ms). The four

Training Sets	Test Sets	M-NN	PCA	LDA	Fuzzy Int.
Set A Set C	Set B Set D	73.77	69.63	73.77	**81,5**
Set B Set D	Set A Set C	71.28	63.71	71.41	**84,7**
Set A Set D	Set B Set C	74.74	69.92	73.12	**80,4**
Set B Set C	Set A Set D	71.36	64.28	74.68	**83,6**

Table 1 Results of the 4 classifiers at audio frame level

different classifiers have been tested using one set of the cartoon (set C or D) and the non-cartoon (set A or B) classes for training and the other for testing. Table 1 represents the four different possibilities when combining the different sets.

Since, the material we have recorded was limited to 2 hours we have used this cross-validation approach to validate the results, and to conclude that it all 4 possible combinations the statistical classification results are similar and over the 80% of detection rate of almost all classification methods reviewed in section 2. When comparing the classification results of the three first classifiers (conventional ones), it can be concluded that in almost all cases the LDA-classifier is the one with the best results with recognition rates above the 75%. It should be also remarked that in the case of the LDA classifier only one coefficient (feature corresponding to the eigenvector associated to the maximal energy) is used whereas in PCA and in Mean-NN, 7 coefficients and 8 features (without projection) are used respectively. As preliminary results, we can observe that the classification results using the fuzzy integral overcome the ones of the other classifiers in a 10% approximately.

5. FUTURE WORK

The results presented in the previous section have used only 8 audiovisual which have been empirically selected studying the Covariance matrix of the feature vector. Again, we will have the problem of high dimensional feature space if we try to combine all the video and features. For this reason, our future work will concentrate on determining those features that are considered as more relevant by the fuzzy integral. Then we will be able to use this information for selecting the optimal subset of features that are really important for the classification and eliminate the measures that have only a minor influence in the final decision. This relevance analysis is necessary due to two main aspects: First, reduction of the computational burden of extracting all the features and second, generalization of the classifier for all possible genres. As illustrated in Fig 1, the classifier uses N features for separating each class C_k but with the help of the relevance

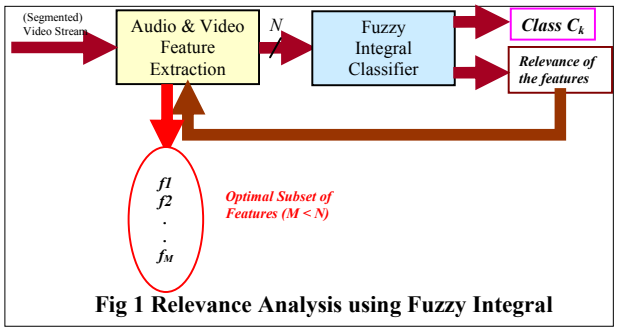

Fig 1 Relevance Analysis using Fuzzy Integral

analysis all the features with small influence in the final classification can be discarded to obtain the optimal subset of M features (with M being smaller than N) that maximizes the classification results and reduces the computational burden on the recognition stage at the same time.

6. CONCLUSIONS

In this paper a novel scheme for detecting cartoons is presented. Preliminary results show a better performance when comparing with other state-of-the art classification methods. Moreover, the fuzzy integral is supposed not only to take a decision for the classification but also to rank all the features and combinations between them as a function of their discriminative power and their capacity of distinguish between the categories. Nevertheless, the proposed scheme still presents some problems that should be overcome. The main problem is the computational cost and the training data required for training the classifier since it has to analyze not only the N audiovisual features but also the combinations between these measures. For this reason, advance computational optimization methods are needed in order to perform the training stage.

7. REFERENCES

[Fischer95] S. Fischer, R. Lienhart and W. Effelsberg, "Automatic recognition of film genres", third ACM International Multimedia Conference and Exhibition, pp. 295-304, 1995.

[Truong00] B. T. Truong, S. Venkatesh and C. Dorai, "Automatic genre identification for content-based video categorization", Proc. 15th Int. Conf. on Pattern Recognition, Vol 4, pp. 230-233, 2000.

[Xu03] L.Q. Xu, Y. Li, "*Video classification using spatial-temporal features and PCA*", *IEEE ICME* V3, pp. 485-8, 2003

[Glasberg05] R. Glasberg, K. Elazouzi and T. Sikora, "Cartoon-Recognition using Visual-Descriptors and a Multilayer-Perceptron" WIAMIS, Montreux, April 13-15, 2005

[Sugeno74] M. Sugeno, "*The Theory of Fuzzy Integrals and Its Applications*", PhD thesis, Tokyo Inst of Technology, 1974.

[Zadeth65] L. A. Zadeh, *Fuzzy sets*, Information Control, (1965), pp. 338–353.

[Soria04] Aureli Soria-Frisch "Soft Data Fusion in Computer Vision", PdD thesis, Fraunhofer Institut, Berlin, May 2004

[Duda01] R.O.Duda, P.E.Hart, D.G.Stork. „Pattern Classification". Second Edition. Wiley