ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, No 5, May 2013

Audio-based Classification of Video Genre Using Multivariate Adaptive Regression Splines

Hnin Ei Latt, Nu War

Abstract— **A large number of researchers are attracted by video genre classification, video contents retrieval and semantics research in video processing and analysis domain. Many researchers try to propose structure or frameworks to classify the video genre that's integrating many algorithms using low and high level features. Features generally include both useful and useless information that are difficult to separate. In this paper, video genre classification is proposed by using only the audio channel. A decomposition model is based on multivariate adaptive regression splines to separate useful and useless components and perform the genre identification is performed on these low-level acoustic features such as MFCC and timbral textual features. Experiments are conducted on a corpus composed from cartoons, sports, news, dahmas and musics on which obtain overall classification rate of 91.83%.**

***Index Terms*— Multivariate Adaptive Regression Splines , Mel Frequency Cepstral Coefficients, Factor Analysis**

I. INTRODUCTION

Today, efficient tools are need for users to crawl the large collection because the available video amount has enlarged significantly on the Internet. While most of the research on video classification has the intent of classifying an entire video, some authors have focused on classifying segments of video such as identifying violent [1] or scary [2] scenes in a movie or distinguishing between different news segments within an entire news broadcast [3]. From this reason, many works are forced on structuring audiovisual databases by content analysis, based on text-based categorization [3]. For the purpose of video classification, features are drawn from three modalities: text, audio, and visual. Most of the proposed approaches rely on image analysis. In [4], many works are motivated by the critical need of efficient tools for structuring audiovisual databases these last years. In [2], they investigate higher level analysis like tracking of audiovisual events. Audio-based approaches were explored by automatic transcription of speech contents, or by low level audio stream analysis. However, these systems generally have poor performances on unexpected linguistic domains and in adverse acoustic conditions.

Acoustic-space characterization is presented by using statistic classifier like gaussian mixture model (GMM), neural nets or support vector machines (SVM) on cepstral domain features [5, 6, 7]. Various kinds of acoustic features have been evaluated in the field of video genre identification. In [8, 7, 5], time-domain audio features are proposed like zero crossing rates or energy distributions. Therefore, low-level approaches present a better robustness to the highly variable and unexpected conditions that may be encountered on videos. In the cepstral domain, one of the main difficulties in genre identification is due to the diversity of the acoustic patterns that may be produced by each video genre. In this paper, this problem is aim to address in the field of identifying video genres by applying multivariate adaptive regression splines. Video genre classification framework is focused on by using an audio-only method.

In the next section an overview of the presented system is provided first. The architecture of the system and the basic underlying concepts are explained. Secondly, the multivariate adaptive regression splines algorithm is described. Finally, the experimental results are also shown.

II. SYSTEM ARCHITECTURE

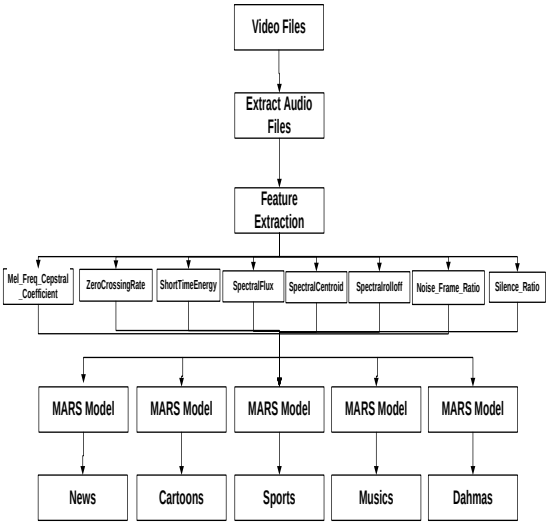


Fig.1. Overview of system architecture

The overall procedure to extract audio file from an video clip has shown in figure.1. Firstly, base audio features are

Manuscript received May, 2013.

Hnin Ei Latt is *with the Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar*

Dr. Nu War *was with the Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar .*

extracted from the audio signal. The MFCC, zero crossing rate, short time energy, spectral flux, spectral centroid, spectral rolloff, noise frame ratio and silence ratio are used as the base audio features in this paper. And then, multivariate adaptive regression splines develop the model for each genre types by using the base audio features set. The next step is an efficient mechanism for classifying genre in the database and measuring their performance. The details of the proposed audio fingerprint are explained in figure.

III. Feature Extraction

Many of the audio-based features are chosen to approximate the human perception of sound. In this frame work uses low-level acoustic features that are both time-domain features and frequency-domain features. The timbral textual features are calculated from the given audio signal. Timbral textual features are those used to differentiate mixture of sounds based on their instrumental compositions when the melody and the pitch components are similar. The use of timbral textural features originates from speech recognition. Extracting timbral features require preprocessing of the sound signals. The signals are divided into statistically stationary frames, usually by applying a window function at fixed intervals. The application of a window function removes the so-called "edge effects." Popular window functions including the Hamming window function. Short-Term Fourier Transform Features: This is a set of features related to timbral textures and is also captured using MFCC. It consists of Spectral Centroid, Spectral Rolloff, Spectral Flux and Low Energy, Zero Crossings and then computes the mean for all five and the variance for all but zero crossings. So, there are a total of nine features. In the time-domain, Zero crossing rate (ZCR) is the number of signal amplitude sign changes in the current frame. Higher frequencies result in higher zero crossing rates. Speech normally has a higher variability of the ZCR than in music. If the loudness and ZCR are both below thresholds, then this frame may represent silence. The silence ratio is the proportion of a frame with amplitude values below some threshold. Speech normally has a higher silence ratio than music. News has a higher silence ratio than commercials. In the frequency-domain, the energy distribution (short time energy) is the signal distribution across frequency components. The frequency centroid, which approximates brightness, is the midpoint of the spectral energy distribution and provides a measure of where the frequency components are concentrated. Normally brightness is higher in music than in speech, whose frequency is normally below 7 kHz. Bandwidth is a measure of the frequency range of a signal. Some types of sounds have more narrow frequency ranges than others. Speech typically has a lower bandwidth than music. The fundamental frequency is the lowest frequency in a sample and approximates pitch, which is a subjective measure. Mel-frequency cepstral coefficients (MFCC) are produced by taking the logarithm of the spectral components and then placing them into bins based upon the Mel frequency scale, which is perception-based.

IV. Multivariate Adaptive Regression Splines

Analyses were performed using multivariate adaptive regression splines, a technique that uses piece-wise linear segments to describe non-linear relationships between audio features and video genre. The theory of multivariate adaptive regression splines (MARS) was developed by Jerome Friedman [9] in 1991. Let z be the dependent response, which can be continuous or binary, and let

$$\mathrm{Y} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n}) \in \mathrm{D} \in \Re^\mathrm{n} \tag{1}$$

be the set of potential predictive covariates. Then the system assume that the data are generated from an unknown "true" model. In case of a continuous response this would be

$$z = f(Y_1, Y_2, \ldots, Y_n) + e \tag{2}$$

The distribution of the error e is member of the exponential family [1]. f is approximated by applying functions, which include interactions of at most second order. That means that use the model

$$f(Y) = g_0 + \sum_{j1} g_{j1}(Y_{j1}) + \sum_{j1<j2} g_{j1,j2}(Y_{j1}, Y_{j2}) + e \tag{3}$$

whereas with error variance . Linear splines and their tensor products are used to model the function g(.). A one-dimensional spline can be written as

$$g(y) = b_{-1} + b_0 y + \sum_{k=1}^{K} b_k (y - t_k)_+ \tag{4}$$

and the knot tk in the range of the observed values of Y. For this reason the function g is situated in a linear space with the K + 2 basis functions. Thus the following model results:

$$g_0 = \beta_0 \quad , \quad g_{j_1}(Y_{j_1}) = \sum_{i=1}^{M} \beta_i^{j_i} B_i^{j_i}(Y_{j_i}) \tag{5}$$

$$g_{j_1,j_2}(Y_{j_1}, Y_{j_2}) = \sum_{i=1}^{M} \beta_i^{j_1 j_2} B_i^{j_1 j_2}(Y_{j_1}, Y_{j_2}) \tag{6}$$

because the interaction gj1,j2 is modeled by means of tensor product splines as

$$g_{12}(y_1, y_2) = g_1(y_1) \times g_2(y_2) \tag{7}$$

The M represent the number of basis functions in the model and the Bs represent spline basis functions as described above and the βs are coefficients. In this approach the coefficients are estimated by using the Least Squares method. Now the coefficient matrix can be written as

$$\hat{\beta} = (Y^{*T} Y^*)^{-1} Y^{*T} Z. \tag{8}$$

Y∗ is represented as the design matrix of the selected basis functions, and Z represents the response vector. Instead of Yj,

MARS uses a collection of new predictors in the form of piecewise linear basis functions are as

$$\{(Y_j - t)_+, (t - Y_j)_+\}, \qquad j = 1, \ldots, n, \qquad t \in \{y_{1j}, .., y_{Nj}\}$$

After that, the generalized cross-validation criterion is used to measure the degree of fit or lack of accuracy of the model :

$$GCV(M) = \frac{\frac{1}{N}\sum_{i=1}^{N}[z_i - \hat{f}_M(y_i)]^2}{[1-\frac{d.M}{N}]^2} \quad (9)$$

whereas denotes the fitted values of the current MARS model and d denotes the penalizing parameter. The numerator is the common residual sum of squares, which is penalized by the denominator, which accounts for the increasing variance in the case of increasing model complexity. A smaller d generates a larger model with more basis functions, a larger d creates a smaller model with less basis functions.

According to the table.1, Forward Process Stage is that the stepwise addition process basis functions are added until the maximal allowed model size is reached. The largest model generally overfits the data. Then Backward pruning Stage -the stepwise deletion process- is that all 'unnecessary' basis functions are removed again until a final model is obtained which is best considering the GCV that is the one with the minimum GCV. In the first step of the addition process a constant model is fitted. Subsequently the number of candidate basis functions depends on the number of possible knots per predictor variable. To keep the procedure fast, the results robust the number of possible knots per predictor and also the possible candidates per step are limited. To determine the number of potential knots of a specific covariate an order statistic is computed and a subset of it is then chosen as potential knots. Commonly these are about 20 knots per predictor, at most every third value is chosen yet. In the first iteration – after the fit of the constant model – a linear basis function on one of the predictor variables is fitted. The second iteration is accounted for both linear basis functions on another covariate and basis functions with knots of the covariate already in the model.

The model to choose in every step during the forward process is the one out of all possible models which minimizes the GCV. In the backward process one basis function is deleted per step and the GCV is computed for the reduced model. The model which yields the smallest increase of GCV becomes the new one.

TABLE.1. ALGORITHM

Forward Process Stage:

- Keeping coefficients the same for variable existed in the current model,
- Update basis functions with the updated side knots
- Add the new basis functions to the model and add the reflected partner
- Select a new basis function pair that produces the largest decrease in training error.
- Repeat the whole process until some termination condition is met:
  - if error is too small or
  - if the number of model's coefficients in the next iteration is expected to be less than number of input variables.

Backward Pruning Stage:

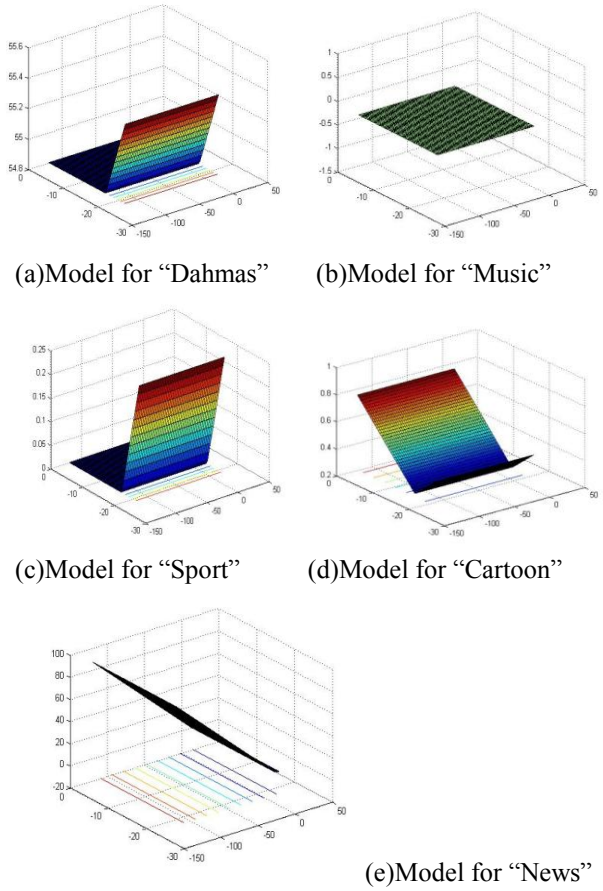
- Find the subset which gives the lowest Cross Validation error, or GCV.
- Delete one basis function per step and reduce model
- Yield the smallest increase of GCV become the new one

V. EXPERIMENTAL RESULTS

The training data set includes the information about a test related with news ,cartoon, sport, music, dama video including 860 observations. Among of them, 140, 235, 176, 113, and 196 clips are 'News', 'Dahma', 'Cartoon', 'Sport' and 'Music' respectively. The data of the dependent variable are binary, and it investigated "whether the sport or not". Here, a 1 is interpreted as "tested positive". This data sample reveals 20 explanatory variables which are given as:Y1 toY13: MFCC, Y14: Zero Crossing Rate, Y15: Short Time Energy, Y16: Spectral Flux, Y17: Spectral Centroid, Y18: Spectral rolloff, Y19: Noise Frame Ratio, and Y20: Silence Ratio.

Using this data set, the proposed framework built the model for each five genre types as show in Figure 2. In model building stage, the parameters is set that maximum number of basis functions is 21 and maximum number of interactions is 2, interpolation type is piecewise-linear, penalty per knot is 3, the best value can also be found using 5-fold Cross-Validation. Larger values will lead to fewer knots being placed

.



(a)Model for "Dahmas" (b)Model for "Music"

(c)Model for "Sport" (d)Model for "Cartoon"

(e)Model for "News"

Fig. 2. MARS Models for each genre

The variable selection results using each MARS model can be summarized in Table 2. It is observed that MFCC3 and MFCC12 do play important roles in deciding the MARS Dahma models. For the MARS Music models, MFCC3 and Noise Frame Rate do play major roles. For the MARS Sports models, Noise Frame Rate, Silence Ratio and Spectral Flux

are more important. Short time energy and MFCC3 are important variable to decide the Cartoon Model. Noise Frame Ratio, Short Time Energy, MFCC3 variables are more important in deciding the New models. From these results, MFCC3, MFCC12, Short Time Energy, Noise Frame Rate, Silance Ratio and Spectral Flux are more useful features than other features.

TABLE 1. DECOMPOSITION ANALYSIS OF EACH MODEL

Gener	Function	STD	GCV	basis	params	Variables
Dahma	1	0.261	0.144	1	2.5	**3**
	2	0.137	0.080	2	5.0	12
	3	0.060	0.056	2	5.0	20
	4	0.050	0.043	1	2.5	2,20
	5	0.113	0.058	2	5.0	3,12
Music	1	3.953	20.21	3	7.5	**3**
	2	0.212	0.074	1	2.5	19
	3	0.038	0.016	1	2.5	3,5
	4	0.085	0.023	1	2.5	3,15
	5	0.050	0.017	2	5.0	3,17
	6	3.840	22.086	3	7.5	3,19
Sports	1	0.087	0.033	1	2.5	2
	2	0.325	0.235	1	2.5	20
	3	0.058	0.027	2	5.0	2,20
	4	0.043	0.025	2	5.0	3,19
	5	0.100	0.038	1	2.5	3,20
	6	0.061	0.028	1	2.5	14,20
	7	0.040	0.025	1	2.5	15,20
	8	0.146	0.051	1	2.5	16,20
	9	0.186	0.082	3	7.5	19,20
Cartoon	1	0.347	0.315	2	5.0	3
	2	0.081	0.060	1	2.5	14
	3	0.351	0.542	2	5.0	15
	4	0.104	0.061	1	2.5	17
News	5	0.358	0.467	2	5.0	19
	6	0.077	0.059	2	5.0	2,15
	7	0.378	0.414	2	5.0	3,15
	8	0.091	0.057	1	2.5	3,19
	9	0.100	0.060	1	2.5	5,17
	10	0.047	0.050	1	2.5	8,14

For this proposed system, a test set of 240 clips are created to test the accuracy rate. Classification accuracy rate is 97 percent for 'Music' genre which is best performance than other genre. It can be seen with matlab program for our own training database that includes 860 videos. As the point of accuracy role, these algorithm has been proofed with true positive rate, true negative rate, false positive rate, and false negative with the own 240 videos database in matlab. According to the Table 2, true positive rate are 48 percent and 56 percent for 'News' and 'Sport' respectively. False positive rate are also reported 54 percent and, 43 percent for 'News' and 'Sport' respectively. True negative rate are shown over 93 percent for each genre. From this table, the proposed approach is not optimized for the 'Sport' and 'News' types.

TABLE.2 CLASSIFICATION

Gener	True Positive	True Negative	False Positive	False Negative	Accuracy
Dahma 49	0.8776 (43)	0.9424 (180)	0.1633 (8)	0.0471 (9)	0.9292 (240)
Music 52	0.9615 (50)	0.9734 (183)	0.0385 (2)	0.0266 (5)	0.9708 (240)
Sport 30	0.5667 (17)	0.9762 (205)	0.4333 (13)	0.0238 (5)	0.9250 (240)
Cartoon 47	0.9333 (42)	0.9333 (182)	0.1111 (5)	0.0564 (11)	0.9333 (240)
News	0.4844 (31)	0.9602 (169)	0.5469 (35)	0.0284 (5)	0.8333 (240)

VI. CONCLUSION

In this paper, Multivariate Adaptive Regression Splines is presented for automatic video genre classification using only audio features. The experiments have indicated that it produce the classification rate which were 91%. As our studies mainly use demographic variables as independent variables, future studies may aim at collecting more important variables to improve the classification accuracies. From the experimental results, MFCC3, MFCC12, Short Time Energy, Noise Frame Rate, Silence Ratio and Spectral Flux are more useful features than other features. The experimental evaluation of this proposed system confirms the good performance of video classification system except sport and new but it is still reasonable results for both types. Further experiments on larger volume of audio, audio-visual features will be tested in this framework. Integrating genetic algorithms and/or grey theory, with neural networks and/or support vector machines are possible research directions in further improving the classification accuracies.

ACKNOWLEDGMENT

My Sincere thanks to my supervisor Dr. Nu War, for providing me an opportunity to do my research work. I express my thanks to my Institution namely University of Technology (Yatanarpon Cyber City) for providing me with a good environment and facilities like Internet, books, computers and all that as my source to complete this research work. My heart-felt thanks to my family, friends and colleagues who have helped me for the completion of this work.

REFERENCES

[1] J. Nam, M. Alghoniemy, and A. H. Tewfik, "Audio-visual content-based violent scene characterization," in International Conference on Image Processing (ICIP '98), vol. 1, 1998, pp. 353–357.

[2] S. Moncriefi, S. Venkatesh, and C. Dorai, "Horror film genre typing and scene labeling via audio analysis," in Multimedia and Expo, 2003, 2003.

[3] W. Zhu, C. Toklu, and S.-P. Liou, "Automatic news video segmentation and categorization based on closed-captioned text," in Multimedia and Expo, ICME, 2001. pp. 829–832

[4] D. Brezeale and D. J. Cook, "Automatic video classification : A survey of the literature," in Systems, Man, and Cybernetics, 2008.
[5] M. Roach, L.-Q. Xu, and J. Mason, "Classification of non-edited broadcast video using holistic low-level features," in (IWDC'2002), 2002.
[6] R. Jasinschi and J. Louie, "Automatic tv program genre classification based on audio patterns," in Euromicro Conference, 2001, 2001.
[7] L.-Q. Xu and Y. Li, "Video classification using spatial-temporal features and pca," in Multimedia and Expo, (ICME '03), 2003.
[8] M. Roach and J. Mason, "Classification of video genre using audio," in European Conference on Speech Communication and Technology, 2001.
[9] Friedman, J.H., "Multivariate adaptive regression splines". Ann. Stat. 19, 1–141 (with discussion) 1991.

First Author Hnin Ei Latt has completed Master of Engineering (Information Technology) (M.E-IT) from West Yangon Technology University (WYTU). Currently, she is a PhD candidate from University of Technology (Yatanarpon Cyber City). She is working as a assistant- lecturer in Technology University (Myeik) and her research areas are videos from the huge amount of video collection..