Movie Genre Classification Using SVM with Audio and Video Features

Conference Paper · December 2012

DOI: 10.1007/978-3-642-35236-2_1

Yin-Fu Huang
National Yunlin University of Science and Technology
161 PUBLICATIONS **1,364** CITATIONS

Movie Genre Classification Using SVM with Audio and Video Features

Yin-Fu Huang and Shih-Hao Wang

Department of Computer Science and Information Engineering
National Yunlin University of Science and Technology
{huangyf,g9817731}@yuntech.edu.tw

Abstract. In this paper, we propose a movie genre classification system using a meta-heuristic optimization algorithm called Self-Adaptive Harmony Search (i.e., SAHS) to select local features for corresponding movie genres. Then, each one-against-one Support Vector Machine (i.e., SVM) classifier is fed with the corresponding local feature set and the majority voting method is used to determine the prediction of each movie. Totally, we extract 277 features from each movie trailer, including visual and audio features. However, no more than 25 features are used to discriminate each pair of movie genres. The experimental results show that the overall accuracy reaches 91.9%, and this demonstrates more precise features can be selected for each pair of genres to get better classification results.

Keywords: Movie genre classification, feature selection, harmony search algorithm, multimedia data mining.

1 Introduction

In recent years, films have become a large portion of the entertainment industry. Every year, there are about 4500 films released around the world and these films approximate 9000 hours of video length [20]. Till now, movie genre classification is still done by man power and has no standardization. In terms of data mining, we would like to explore the hidden differences between movie genres by analyzing the information and/or features in videos.

In order to gather up the enough knowledge for classification, a generated feature set commonly contains the abundant information with some probably redundant features. For solving this problem, dimensionality reduction techniques are frequently employed and they can be classified into two approaches. One kind is to transform the matrix of a feature set from a high-dimensional space to a lower-dimensional space through the linear combination of matrix using the techniques such as principal component analysis (PCA) [9], non-negative multi-linear principal component analysis (NMPCA) [17], non-negative tensor factorization (NTF) [16], et al. Another kind is called the feature selection which finds an optimum subset from the original feature set using the search algorithms such as genetic algorithm (GA) [10], ant colony

R. Huang et al. (Eds.): AMT 2012, LNCS 7669, pp. 1–10, 2012.

optimum (ACO) [3], harmony search (HS) [4], et al. Both these two approaches can effectively reduce the dimensions of a feature set. In our work, we extensively collect useful features and aim to evaluate which features are more relevant. Here, a feature selection approach called the self-adaptive harmony search (SAHS) [7] algorithm is adopted to obtain a better feature subset. In general, feature selection approaches are to select a global feature set for all movie genres. However, in order to achieve a global optimization, we adopt a local selection strategy based on each pair of movie genres since it can derive a more relevant local feature set than a global selection strategy.

For the prediction, an SVM classifier is adopted since, in general, it presents better performances than other classifiers [6, 8, 18] while a kernel function and parameters are appropriately chosen. In this paper, we match each local feature set with an SVM classifier and use the majority voting method to determine the prediction of each movie. The experimental results verify that more precise features can be selected for each pair of genres to get better classification results.

The remainder of the paper is organized as follows. In Section 2, we propose the system architecture and describe the visual and audio features used to discriminate movie genres. In Section 3, the SAHS algorithm and correlation measuring method are proposed to derive an optimum feature subset from the original feature set. In Section 4, the experimental results based on different local features for each pair of genres are presented. Finally, we make conclusions in Section 5.

2 System Overviews

In this section, we proposed an automatic movie classification system. As we know, features extracted from movies play an important role in movie classification. If relevant features are selected, movies would be easily distinguished based on them. Here, a meta-heuristic optimization algorithm called Self-Adaptive Harmony Search (i.e., SAHS) algorithm [7] is employed to select features for corresponding movie genres. The feature selection mechanism is applied on each one-against-one classifier (or SVM) to enable ambiguous genres to be classified more precisely. As illustrated in Fig. 1, the system architecture consists of two parts: training phase and test phase.

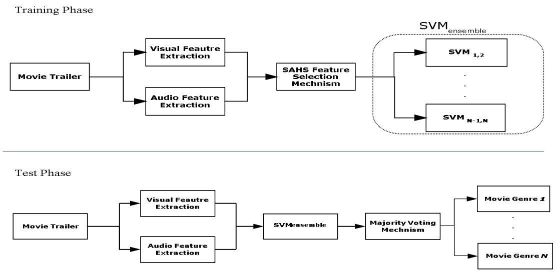

Fig. 1. System architecture

In the training phase, retrieving the visual and audio features from the training set is the first step. Visual features can be classified into temporal and spatial features. Temporal features are the ones relevant to video timing, which are the results of shot boundary detection, such as shot number, average shot number, total frame number, and average shot length. On the other hand, spatial features are more well-known than temporal features; they are MPEG-7 feature descriptors [2, 13, 14] all extracted from key-frames directly. For audio features, spectrum, compactness, Zero Crossing Rate (ZCR), Root Mean Square (RMS), Linear Prediction Coefficients (LPC), and MFCC are extracted in the system. For these collected features, the feature selection mechanism is used to pick up relevant features for each pair of genres, and then produces their optimum feature set (i.e., a local feature set). Thus, for N-genre videos to be classified, $C\binom{N}{2}$ local feature sets would be generated in the mechanism. Next, each one-against-one SVM is trained by the corresponding local feature set. Finally, these trained SVMs are combined together to form an SVM ensemble model used in the test phase.

For the test phase, we also retrieve the visual and audio features from the test set first. Then, they are fed into the SVM ensemble model. Finally, through the majority voting on $C\binom{N}{2}$ local predictions, the genre of a test sample could be determined.

2.1 Visual Features

In general, visual signals can be measured in two different ways: temporal and spatial features. To extract these visual features as shown in Fig. 2, a video should be preprocessed using shot boundary detection [12] to redefine the video as a series of video shots from which temporal features can be got. Then, key-frames are selected from each video shot, and spatial features can be extracted from these key-frames.

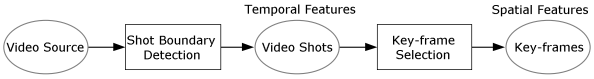

Fig. 2. Visual feature extraction

Temporal Features. In general, the rhythms of each film categories are totally different from others. For example, an action movie with fighting scenes, gun wars, car crashes, and explosion scenes always has a faster tempo. Thus, the temporal features as illustrated in Table 1 are viewed as the related ones in movie classification work.

Table 1. Temporal features

No.	Feature description	Dim.	Overall statistics	Total number
1	Shot number	1	1	1
2	Average shot number	1	1	1
3	Total frame numbers	1	1	1
4	Average shot length	1	1	1

Spatial Features. The spatial features defined by *MPEG-7 feature descriptors* [14] are widely used in existing movie classification work. MPEG-7 is formally known as the multimedia content description interface in the ISO/IEC 15938 standard developed by Moving Picture Experts Group (i.e., MPEG). It addresses multimedia contents with various modalities including image, video, audio, speech, graphics, and their combinations. The ultimate goal and objective of MPEG-7 is to provide interoperability among systems and applications used in generation management, distribution, and consumption of audio-visual content descriptions. As illustrated in Table 2, we only use motion vector, average color histogram, and average lighting key defined in Motion Activity and Color Layout descriptors as spatial features.

Table 2. Spatial features

No.	Feature description	Dim.	Overall statistics	Total number
5~10	Motion vector	1	6	6
11~74	Average color histogram	1	64	64
75	Average lighting key	1	1	1

2.2 Audio Features

In general, audio signals can be measured in two different ways: time domain and frequency domain. For the audio features in the time domain, the samples of audio signals are directly processed along time so that we can observe some characteristics about the amplitude such as intensity and rhythm. For the frequency domain, each amplitude sample is transformed from the time domain to a corresponding frequency band in a spectrum through the discrete Fourier transform (i.e., DFT). More detailed characteristics about acoustics such as timbre are presented by estimating the spectrum. All audio features are extracted using the well-known software “jAudio” [15].

Intensity. In acoustics, intensity as illustrated in Table 3 is also called loudness, volume, or energy. It is the most obvious feature and means the loud volume heard by human perception. It is commonly measured by the amplitude within a frame in the time domain. In general, the unit of intensity is represented in decibel (i.e., dB). The Root Mean Square (i.e., RMS) amplitude is a measure of the power of a signal, which can be expressed as follows:

$$P(n) = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} S_n^2(i)} \tag{1}$$

where $S_n(i)$ is the ith sample of the nth audio frame and N is the total sample numbers in the frame. Besides, the fraction of low energy windows is to measure how many windows (or frames) are quiet relative to the others in the audio signal. First, the mean of the RMS amplitude of the last 100 windows is calculated, and the fraction of these 100 windows with RMS amplitude below the mean can be found.

Table 3. Intensity features

No.	Feature description	Dim.	Overall statistics	Total number
76~83	RMS amplitude	1	8	8
84~91	Fraction of low energy windows	1	8	8

Timbre. Timbre is an audio feature used to distinguish the sounds which have the same intensity and pitch. Different timbre is presented in different structures of amplitude spectrum on each or all frequency bands. By estimating the spectrum of audio signals, we can derive some timbre characteristics. In our work, we consider total 178 timbre features as illustrated in Table 4. Spectral centroid estimates the centroid frequency of spectrum; spectral flux estimates the distance of spectrum between adjacent frames; spectral variability estimates the variation degree of the neighboring peaks of spectrum; zero crossing counts the number of the signals crossing the zero line; compactness is closely related to harmonic spectral smoothness but is estimated through amplitudes; MFCC offers a description of the spectral shape based on Mel-frequency; finally linear prediction coefficients are calculated using autocorrelation and Levinson-Durbin recursion [11].

Table 4. Timbre features

No.	Feature description	Dim.	Overall statistics	Total number
92~99	Spectral centroid	1	8	8
100~107	Spectral flux	1	8	8
108~115	Spectral variability	1	8	8
116~123	Zero crossing	1	8	8
124~131	Compactness	1	8	8
132~209	MFCC	13	6	78
210~269	LPC	10	6	60

Rhythm. Rhythm may be generally defined as a "movement marked by the regulated succession of strong and weak elements, or of opposite or different conditions" [23]. In our viewpoint, rhythm is a time-related characteristic in sounds, and consists of beats and tempos. A beat is related to notes and commonly produced by striking or hitting. It is usually measured by the peaks of amplitude. As illustrated in Table 5, the strongest beat is the strongest bin in the beat histogram.

Table 5. Rhythm features

No.	Feature description	Dim.	Overall statistics	Total number
270~277	Strongest beat	1	8	8

3 Feature Selection Using SAHS

The purpose of feature selection is to select the most relevant features to facilitate the classification, using subset selection algorithms. These selection algorithms could be categorized into three different approaches: wrappers, filters, and embedded. The wrapper approach utilizes the performance of a target classifier to evaluate feature subsets; however it is computationally heavy since a target classifier must be iteratively trained on each feature subset. The filter approach evaluates the performance of feature subsets through an independent filter model before a target classifier is applied. The embedded approach must utilize the specific classification algorithm containing the evaluation body inside. Here, we use the filter approach to select the most relevant features since its computational loading is acceptable and the selected classification algorithm could be general. In our work, the feature selection model consists of two parts: Self-Adaptive Harmony Search (i.e., SAHS) algorithm [7] and relative correlations, as illustrated in Fig. 3. Once the original feature set is given, the SAHS algorithm starts to iteratively search the better solution that would be evaluated later by the relative correlations. Finally, the best solution will be output as the final feature subset.

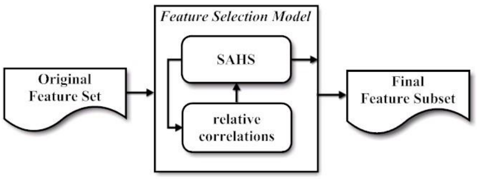

Fig. 3. Feature selection model

3.1 Relative Correlations

When each new harmony symbolizing a selected feature subset is generated from the SAHS algorithm, the relative correlations are used to evaluate the performance of the selected feature subset. The correlations of the selected feature subset are conducted in two phases; one is the intra-correlation evaluating the mutual correlation between features within the subset, and another is the inter-correlation comparing each feature inside the subset with the corresponding class. If a subset has better performance, it must possess the property of lower intra-correlation and higher inter-correlation. The lower intra-correlation means the features within the subset are relevant, whereas the higher inter-correlation means each feature within the subset is discriminative for the corresponding class.

Here, the well-known measuring formula called mutual information [19] is adopted, which evaluates the degree of the mutual dependence between two variables. The definition for discrete random variables is shown as follows.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right) \tag{2}$$

where $p(x,y)$ is the joint probability density of X and Y, and $p_1(x)$ and $p_2(y)$ are the marginal probability distribution functions of X and Y, respectively. The value of $I(X;Y)$ is nonnegative; if $I(X;Y)$ is zero, it means X and Y are independent; otherwise, a high value indicates a strong dependence between X and Y.

The intra-correlation within the feature subset is shown as follows.

$$RI(S) = \frac{1}{C(|S|,2)} \sum_{i=1}^{|S|} \sum_{j=i+1}^{|S|} I(x_i; x_j) \tag{3}$$

where C(|S|,2) is the number of 2-combination on the cardinality of feature subset S. The overall correlation within subset S is divided by C(|S|,2) to present the average correlation between features within subset S. The inter-correlation between the feature subset and the corresponding class is shown as follows.

$$RT(S,y) = \frac{1}{|S|} \sum_{i=1}^{|S|} I(x_i; y) \tag{4}$$

where |S| is the cardinality of feature subset S, and y is the output class. The overall correlation is divided by the cardinality to derive the average correlation between features and the corresponding class.

Finally, the relative overall correlation combining both the intra-correlation and inter-correlation [5] is shown as follows.

$$RC(S,y) = \frac{k \times RT(S,y)}{\sqrt{k + k \times (k-1) \times RI(S)}} \tag{5}$$

where k is the cardinality of feature subset S. It not only indicates a higher RC value has a better selected feature subset (i.e., it has lower intra-correlation and higher inter-correlation), but also balances the effect of average correlation on different cardinalities of candidate feature subsets. As a result, the RC is chosen as the objective function in the feature selection model.

4 Experimental Results

In our work, we collect 223 movie trailers from the Apple Movie Trailers website [21], including 35 for action, 14 for animation, 46 for comedy, 28 for documentary, 67 for drama, 7 for musical, and 26 for thriller. The genres of these 223 movies are defined by the Internet Movie Database (IMDb) [22]. In the SAHS algorithm, the parameters HMS and HMCR are set with 50 and 0.99 to present the optimum performance. Next, we adopt the well-known LIBSVM developed by Chang [1] as an SVM classifier. The kernel function used here is RBF (i.e., Radial Basis Function) since it is more accurate and effective than the other kernel ones. The parameters γ and C are determined by the optimum performance of 6X6 combinations between

$[2^{-4},\ldots,2^{1}]$ and $[2^{-2},\ldots,2^{3}]$. Moreover, the feature vector is normalized as the range [-1, 1]. For the classification results, a confusion matrix is employed to present the entire statistic on the correct and false predictions after the cross validation. Precision, recall, and accuracy are estimated as the performance evaluation and showed as follows.

$$\begin{aligned}\Pr ecision &= \frac{N_C}{N_C + N_F}\\ \operatorname{Re} call &= \frac{N_C}{N_C + N_M}\\ Accuracy &= \frac{total_C}{total_M}\end{aligned} \tag{6}$$

4.1 Classification Results Using the Feature Selection

The classification results using the local feature selection are presented here. Each feature set selected based on the local selection strategy is for each pair of all seven genres. In the experiments, for seven genres, 21 one-against-one local feature sets are generated for classification. The number of features in each local feature set is much less than the original 277 features, as illustrated in Table 6. We find that Color Histogram, Motion Vector, MFCC, and LPC are four critical feature types appearing in most pairs. For the classification results illustrated in Table 7, except the recall in

Table 6. Number of features in each local feature set

Genre	action	animation	comedy	documentary	drama	musical	thriller
action	*	13	11	21	10	5	3
animation	-	*	15	12	19	2	11
comedy	-	-	*	11	14	3	15
documentary	-	-	-	*	2	3	10
drama	-	-	-	-	*	2	11
musical	-	-	-	-	-	*	5
thriller	-	-	-	-	-	-	*

Table 7. Confusion matrix by the local selection strategy

Predicted Actual	action	animation	comedy	documentary	drama	musical	thrill	*Recall (%)*
action	35	0	0	0	0	0	0	100
animation	0	13	0	0	1	0	0	92.9
comedy	0	0	45	0	1	0	0	97.8
documentary	0	0	1	23	4	0	0	82.1
drama	0	0	0	0	66	0	1	98.5
musical	0	0	0	0	1	6	0	85.7
thriller	1	0	0	0	8	0	17	65.4
Precision (%)	97.2	100	97.8	100	81.5	100	94.4	

thrill, we get good precision and recall in each genre. We find that it is usual to identify thrill movies as drama movies, even when we use both visual and audio features in genre classification. However, the overall accuracy is around 91.9%, and this demonstrates that, without considering the other genres, more precise features can be selected for each pair of genres to get better classification results.

4.2 Comparisons among All Methods

Finally, the qualitative comparisons between our methods and the other start-of-art methods are illustrated in Table 8. Although we extract two feature types (i.e., visual and audio) including 277 features from movie trailers, more movie genres could be classified more precisely using the local feature selection. In fact, no more than 25 features as shown in Table 6 are used to discriminate each pair of movie genres.

Table 8. Comparisons among all methods

	Z. Rasheed et al. [18]	H.Y. Huang et al. [6]	S.K. Jain et al. [8]	Ours
No. of genres	4	3	5	7
Feature types	Visual	Visual	Visual-Audio	Visual-Audio
Feature dim.	4	4	21	277
Feature selection	N	N	N	Y
Classifier	Mean Shift Classification	2-Layer Neural Network	Neural Network	SVMs
Accuracy	83%	80.2%	87.5%	91.9%

5 Conclusions

In this paper, a movie genre classification system is proposed. Totally, 277 visual and audio features are extracted from movie trailers. By employing the SAHS (self-adaptive harmony search) algorithm on these 277 features, the feature selection model can effectively find the optimum feature subsets for corresponding movie genres. From the results of the SAHS algorithm, we find audio features are more relevant than visual features in discriminating movie genres. Finally, the experimental results show that the overall accuracy reaches 91.9%, and this demonstrates more precise features can be selected for each pair of genres to get better classification results. In the future, we hope high-level features or semantics can be explored to classify movie genres more precisely.

Acknowledgments. This work was supported by National Science Council of R.O.C. under Grant NSC100-2218-E-224-011-MY3.

References

1. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
2. Chang, S.-F., Sikora, T., Puri, A.: Overview of the MPEG-7 standard. IEEE Transactions on Circuits and Systems for Video Technology 11(6), 688–695 (2001)
3. Deriche, M.: Feature selection using ant colony optimization. In: Proc. the 6th International Multi-Conference on Systems, Signals and Devices, pp. 1–4 (2009)
4. Diao, R., Shen, Q.: Two new approaches to feature selection with harmony search. In: Proc. the IEEE International Conference on Fuzzy Systems, pp. 1–7 (2010)
5. Hall, M.A., Smith, L.A.: Practical feature subset selection for machine learning. In: Proc. the 21st Australian Computer Science Conference, pp. 181–191 (1998)
6. Huang, H.Y., Shih, W.S., Hsu, W.H.: Movie classification using visual effect features. In: Proc. the IEEE Workshop on Signal Processing Systems, pp. 295–300 (2007)
7. Huang, Y.-F., Wang, C.-M.: Self-adaptive harmony search algorithm for optimization. Expert Systems with Applications 37(4), 2826–2837 (2010)
8. Jain, S.K., Jadon, R.S.: Movies genres classifier using neural network. In: Proc. the 24th International Symposium on Computer and Information Sciences, pp. 575–580 (2009)
9. Jin, X., Bie, R.: Random forest and PCA for self-organizing maps based automatic music genre discrimination. In: Proc. the International Conference on Data Mining, pp. 414–417 (2006)
10. Lanzi, P.L.: Fast feature selection with genetic algorithms: a filter approach. In: Proc. the International Conference on Evolutionary Computation, pp. 537–540 (1997)
11. Levinson, N.: The Wiener RMS error criterion in filter design and prediction. J. Math. Phys. 25, 261–278 (1947)
12. Lienhart, R.: Comparison of automatic shot boundary detection algorithms. In: Proc. Storage and Retrieval for Image and Video Databases VII. SPIE, vol. 3656, pp. 1–12 (1998)
13. Martinez, J.M., Koenen, R., Pereira, F.: MPEG-7: the generic multimedia content description standard, part 1. IEEE Multimedia 9(2), 78–87 (2002)
14. Martinez, J.M.: MPEG-7 overview (version 10), ISO/IEC JTC1/SC29/WG11 N6828 (2004)
15. McEnnis, D., McKay, C., Fujinaga, I., Depalle, P.: jAudio: a feature extraction library. In: Proc. the 6th International Conference on Music Information Retrieval, pp. 600–603 (2005)
16. Panagakis, Y., Kotropoulos, C.: Music genre classification via topology preserving non-negative tensor factorization and sparse representations. In: Proc. the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 249–252 (2010)
17. Panagakis, Y., Kotropoulos, C., Arce, G.R.: Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification. IEEE Transactions on Audio, Speech, and Language Processing 18(3), 576–588 (2010)
18. Rasheed, Z., Sheikn, Y., Shah, M.: On the use of computable features for film classification. IEEE Transactions on Circuits and System for Video Technology 15(1), 52–64 (2005)
19. Silviu, G.: Information Theory with Applications. McGraw-Hill (1977)
20. Wactlar, H.D.: The challenges of continuous capture, contemporaneous analysis, and customized summarization of video content. In: Proc. the Workshop on Defining a Motion Imagery Research and Development Program, pp. 1–9 (2001)
21. Apple iTunes Movie Trailers Website at `http://trailers.apple.com/trailers/`
22. Internet Movie Database (IMDb) at `http://www.imdb.com/`
23. The Compact Edition of the Oxford English Dictionary II, p. 2537. Oxford University Press (1971)