Harmonic and Instrumental Information Fusion for Musical Genre Classification

Tomás Pérez-García
Department of Software and Computing Systems
University of Alicante
Alicante, Spain
tperez@dlsi.ua.es

Carlos Pérez-Sancho
Department of Software and Computing Systems
University of Alicante
Alicante, Spain
cperez@dlsi.ua.es

José M. Iñesta
Department of Software and Computing Systems
University of Alicante
Alicante, Spain
inesta@dlsi.ua.es

ABSTRACT

This paper presents a musical genre classification system based on the combination of two kinds of information of very different nature: the instrumentation information contained in a MIDI file (metadata) and the chords that provide the harmonic structure of the musical score stored in that file (content). The fusion of these two information sources gives a single feature vector that represents the file and to which classification techniques usually utilized for text categorization tasks are applied. The classification task is performed under a probabilistic approach that has improved the results previously obtained for the same data using the instrumental or the chord information independently.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications

General Terms

Experimentation

Keywords

Genre classification, multimodality

1. INTRODUCTION

Multimodality is gaining more importance in pattern recognition research. Obtaining descriptive features from an object from different information sources permits one to perform a deeper and more informative description of it.

Some problems related to music information extraction and retrieval adapt well to this scenario. For example, the classification of a music score filed in a symbolic format into a genre selected from a catalogue [9]. We are going to pose this problem in this paper extracting and combining descriptors of different nature from the container file.

A number of papers can be found in the literature where pattern recognition is based on multimodal information. In [8] the authors explain how multimodality in human interaction and multimedia information processing can help to improve the performance in different pattern recognition tasks, like manuscript text processing or gesture recognition from image sequences.

In [2] the authors consider a video sequence as a multimodal information source, obtaining features of different nature from speech, audio, text, shapes, or colors. This approach works under an *early* scheme where features are combined in a compact representation for a single decision.

Other approaches use a *late* scheme where various classifiers are utilized for the different information sources and are then combined into a decision. For example, in [3] a multiple classifier system for OCR is presented, based on hidden Markov models that provide individual decisions. The combination of them is performed with a voting system.

Using instrumental information for genre classification in a multimodal context is not new. In [5] a method is described for extracting high-level features from MIDI files. Those features are utilized either individually or combining them hierarchically in order to perform the classification. The classification was performed using neural networks and k-NN methods. The authors reported the best classification performance using the features related to the analysis of the instruments utilized in the sequence. Also, in [4] the authors have addressed this problem under a multimodal approach combining audio and lyrics features.

In the present work, we are going to study how the information from two different sources can be useful in musical genre classification, combining the instrumentation information contained in a MIDI file metadata with the chords provided by the harmonic structure of the music sequence. Using an early combination approach, the features derived from both sources will give a single vector that will be the input to a classifier after a feature selection procedure.

Each feature codifies the presence or absence of a given item (an instrument or a chord) and there will be a probability of each feature associated to each class, depending on the frequencies found in the training set for the items in the classes. The decision will be taken combining these probabilities through a naïve Bayes classifier.

2. METHODOLOGY

In [7] text categorization technologies were used for musical genre recognition from the harmonic information contained in a number of works stored in digital scores. In the present work the same method will be used to design a naïve

MML'10, October 25, 2010, Firenze, Italy.

Bayes classifier to infer the genre of a music work from the chords that harmonize the sequence and the instrumentation metadata provided with the MIDI file. Both subsets of features will be combined in a single representation.

A set of classes $C = \{c_1, c_2, .., c_{|C|}\}$ will be available and a training set of files $X = \{x_1, x_2, .., x_{|X|}\}$ containing the music works labeled with the genre they belong to.

Each file in the dataset is represented as a vector $\mathbf{x} \in \{0,1\}^{H+I}$, where each component $x_i \in \{0,1\}$ codes the presence or absence of the i-th feature. H denotes the number of chords in the dictionary of possible harmonic combinations considered, $H = 312$ different chords in this work (see [7] for more details), and I is the number of possible instruments that, assuming the General MIDI standard for the sequence, will be 128 instruments plus 3 percussion sets. Therefore, $I = 131$.

A given test file $\mathbf{x}$ is assigned to the class c_j with the maximum a posteriory probability, computed as:

$$P(c_j|\mathbf{x}) = \frac{P(c_j)P(\mathbf{x}|c_j)}{P(\mathbf{x})} \quad , \tag{1}$$

where $P(c_j)$ is the a priori probability of the class c_j computed in this case as $P(c_j) = 1/|C|$, assuming that all the genres are equally probable. $P(\mathbf{x}) = \sum_{j=1}^{|C|} P(c_j)P(\mathbf{x}|c_j)$ is a normalization factor, and $\mathrm{P}(\mathbf{x}|c_j)$ is the probability of $\mathbf{x}$ being generated by the class c_j following a Bernoulli multivariate distribution over the instruments and chords of class c_j, learnt from the training set:

$$P(\mathbf{x}|c_j) = \prod_{i=1}^{D} x_i P(t_i|c_j) + (1 - x_i)(1 - P(t_i|c_j)) \tag{2}$$

where $\mathrm{P}(t_i \mid c_j)$ are the probabilities of each instrument or chord t_i for each class c_j, that can be computed by counting the number of occurrences of each item in each class c_j.

$$P(t_i|c_j) = \frac{1 + M_{ij}}{2 + M_j} \tag{3}$$

where M_{ij} is the number of files of class c_j where the instrument or chord t_i appears, and M_j is the total number of songs of class c_j. This equation permits to avoid zero valued probabilities when an item appears in the test song that was not observed during the training phase.

This method uses a representation of the music works as a vector of symbols with associated probabilities. A common practice in text classification is to reduce the dimensionality of these vectors before training the classifier, through a feature selection procedure. This process avoids model overtraining when the number of samples available is limited and there a large number of features. It also speeds up the performance of the system.

In order to select the features that contribute the most to class discrimination, a feature ranking has been established based on the *Average Mutual Information* (AMI) [1], that provides a measure of how much information about a class is able to provide a single feature. Informally, one can consider a given feature as being very informative if it is very frequent in a class while being seldom utilized in the others.

AMI is computed between the class of a song and the presence or absence of a feature in it. C is defined as a random variable over all classes and F_i as a random variable over the presence or absence of a feature t_i in a file. F_i takes values in $f_i \in \{0,1\}$, where $f_i = 0$ codes the absence of t_i and $f_i = 1$ codes its presence. AMI is computed for all t_i as

$$I(C; F_i) = \sum_{j=1}^{|C|} \sum_{f_i \in \{0,1\}} P(c_j, f_i) log \frac{P(c_j, f_i)}{P(c_j)P(f_i)} \tag{4}$$

where $P(c_j)$ is the number of music works for class c_j divided by the total number of works; $P(t_i)$ is the number of songs that contain that feature divided by the total number of songs; and $P(c_j, f_i)$ is the number of works in class c_j that contains the value f_i for the feature t_i divided by the total number of songs.

3. EXPERIMENTS

In the following experiments, a database made up of 856 music files has been used. It is divided in three musical genres (3-G): academic, jazz, and popular music. A second split of this database divides each genre in three subgenres, resulting in a total of 9 music subgenres (9-G). The number of files eventually used for each genre is displayed in Table 1, along with their corresponding decomposition in subgenres.

Table 1: Number of files per genre and subgenre

Academic	235	**Jazz**	338	**Popular**	283
Baroque	56	Pre-bop	178	Blues	84
Classical	50	Bop	94	Pop	100
Romanticism	129	Bossanova	66	Celtic	99

This dataset is available in MIDI and Band-in-a-Box formats (see [7] for details). The chord sequences were obtained from the Band-in-a-Box files, while the instrumental features were extracted from the MIDI files using the *metamidi* tool, as described in [6].

All the experiments have been performed following a 10-fold cross-validation scheme, using the mean and standard deviation of the 10 subexperiments as the success rate and confidence interval in the results reported below.

3.1 Previous work

3.1.1 Instrument-based classification of genres

In [6], the authors performed genre classification using just the instrumental information extracted from MIDI files. Thus, using the methodology presented in Section 2, the feature vectors are defined as $\mathbf{x} \in \{0,1\}^I$, where each component of the vector $x_i \in \{0,1\}$ represents the presence or absence of a certain instrument t_i. Using this approach, the conditional probability of a MIDI file to belong to a certain genre is based on the probability that the instruments present in the file belong to that same genre.

The results in genre classification using this approach are shown in Table 2. The success rate for this experiment was 93 ± 2, showing that instrumental information is highly useful for music genre classification. The confusion matrix for this experiment is shown in Table 3. In the more complex classification task using the 9-G database, the success rate descended to 68 ± 5.

At the light of these results, it can be concluded that instrumental information is useful in order to perform automatic classification between broad genres. However, when using a finer division of genres, as in the case of the 9-G problem, classification cannot be properly performed due to

Table 2: Success rates in classification using instrumental features.

	% Success
3 genres	93 ± 2
9 genres	68 ± 5

Table 3: Confusion matrix using 3-G database and instrumental features.

	Academic	Jazz	Popular
Academic	228	6	1
Jazz	3	295	40
Popular	5	16	262

the coincidence of instruments in the different subgenres. The fact that the errors are committed mainly inside the same root genre is an indicator that instrumental information alone is not enough for this task.

3.1.2 Harmonic classification of genres

In [7], a text classification approach is used in order to perform genre classification using the harmonic information contained in the songs. In that work, a Naïve Bayes classifier and language modeling with n-grams were used using the same database. As in the previous experiment, when using the Naïve bayes approach each file in the database is represented as a feature vector $\mathbf{x} \in \{0,1\}^H$, where each vector component $x_i \in \{0,1\}$ encodes the presence or absence of a certain chord.

The results obtained in that work are shown in Table 4. As it can be seen, the harmonic information provided by the chords led to a classification rate of 86 ± 4 in the 3-G problem, and a 64 ± 4 in the 9-G problem. Again, most of the errors committed in the 9-G problem occurred between the subgenres within the same root genre. Table 5 shows the confusion matrix for the 3-G problem, using chords and the naïve Bayes classifier. Note that these success rates are significantly lower than those of the previous experiment using instrumental information.

Table 4: Success rates in classification using chords.

	n-grams	naïve Bayes
3 genres	86 ± 4	83 ± 3
9 genres	40 ± 10	64 ± 4

3.2 Harmonic and instrumental information

In this experiment we have used the same methodology and data as in the previous experiments, using both the 3-G and 9-G versions of the dataset. However, for recognizing the musical genre of a piece, the harmonic and instrumental information are fused, generating a unique set of features. Thus, each music file is represented as a vector of harmonic

Table 5: Confusion matrix for the 3-G problem using chords and the naïve Bayes classifier.

	Academic	Jazz	Popular
Academic	183	2	50
Jazz	2	304	32
Popular	32	27	224

and instrumental features $\mathbf{x} \in \{0,1\}^{I+H}$, where each component $x_i \in \{0,1\}$ encodes the presence or absence of a certain feature.

The results for this experiment are shown in Table 6. As it can be seen, a significant improvement has been obtained in the 9-G problem, reaching a 79 ± 3 classification rate. In the 3-G problem, however, although the 95 ± 2 classification rate is also higher than in the two previous experiments, the difference cannot be considered statistically significant compared to the results obtained in the experiments using instrumental features only.

Table 6: Success rates in classification using harmonic and instrumental features.

	% Success
3 genres	95 ± 2
9 genres	79 ± 3

As for the combined feature set, Figure 1 shows the evolution of the classification rate as a function of the number of features used. Note that, from the 443 possible features $(I + H)$, only 300 of them appeared in the dataset used in the experiments. It can be seen that, in the 3-G problem, the highest success rate is obtained when using 1/3 of the total number of features. In the 9-G problem the highest success rate is achieved when using the whole feature set, which can be an indicator that the training dataset is not big enough for this task. Note also that with a subset of around 50 features, the results obtained are similar to those reached with all of them.

We have performed a study of these 50 more relevant features, according to the ranking produced by the AMI feature selection method. We have found that both the harmonic and instrumental features have the same relevance in classification, using approximately 14% of both of them in the 50-features subset. We have also observed a coincidence of a 85% of the selected features in the 3-G and 9-G problems.

The confusion matrices for the 3-G and 9-G problems are shown in Tables 7 and 8 respectively. As it can be seen, although the result obtained in the 9-G problem is significantly better than those obtained in the previous experiments, the majority of the errors are still committed between the subgenres within the same root genre.

Finally, the comparison between all the experiments is shown in Table 9. Here we can see that the union of the two kinds of features improves the results obtained with the individual feature sets.

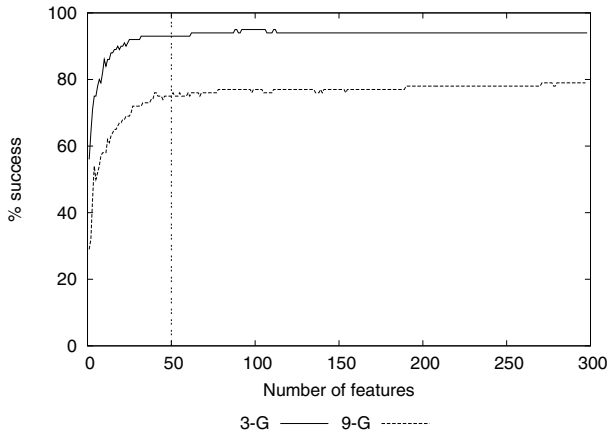

Figure 1: Evolution of classification according to the number of features used.

Table 7: Confusion matrix for the 3-G problem using harmonic and instrumental features.

	Academic	Jazz	Popular
Academic	255	1	9
Jazz	0	328	10
Popular	6	17	262

Table 8: Confusion matrix for the 9-G problem using harmonic and instrumental features.

	bar	cla	rom	pre	bop	bos	cel	blu	pop
bar	32	3	19	0	0	0	1	0	1
clas	6	20	24	0	0	0	0	0	0
rom	8	12	104	0	0	1	1	2	1
pre	0	0	0	152	21	5	0	0	0
bop	0	0	0	32	55	57	0	0	4
bos	0	0	0	2	3	57	0	0	4
cel	0	0	0	0	0	0	99	0	0
blu	0	0	1	1	2	0	0	75	5
pop	0	1	2	6	0	4	3	6	78

Table 9: Success rates in classification using different feature sets.

	Instrumental	Harmonic	Instrumental + harmonic
3 genres	93 ± 2	86 ± 4	95 ± 2
9 genres	68 ± 5	62 ± 3	79 ± 3

4. CONCLUSIONS AND FUTURE WORK

In this work we have shown that the combination of harmonic and instrumental features in musical genre classification provides better results than those obtained in previous experiments using these two kind of features separately. In these experiments we have used a dataset with a two-level taxonomy, with 3 genres in the first level and 9 subgenres in the second. At the first level of the hierarchy, the results obtained using the combined feature set outperformed those using harmonic information alone. However, this improvement was not significant compared to the instrumental feature set alone. On the other hand, a significant improvement has been achieved when classifying between the 9 musical subgenres, reaching a 79 ± 3 classification rate.

A study performed on the features used in classification revealed that both kinds of information, harmonic and instrumental, have the same weight in this task, since there is not a prevalence of neither of them over the other in a selected subset of the 50 more relevant features using a feature selection algorithm. However, in the 9 genres classification task, this feature selection procedure did not provide any advantage, since the highest classification rate was obtained using the whole feature set. We interpret this fact as an indicator that the dataset is not big enough for this task.

As future work, we plan to increase the size of the dataset, and also to obtain additional features from different sources of information, in order to perform musical genre classification from a multimodal point of view.

5. ACKNOWLEDGMENTS

This work has been financed by the projects from the spanish ministry MICINN: TIN2009-14247-C02-02 and Consolider Ingenio 2010 (MIPRCV, CSD2007-00018), both of them partially financed by UE ERDF.

6. REFERENCES

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[2] Z. Liu, Y. Wang, and T. Chen. Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI Signal Processing System*, pp. 61–79, 1998.

[3] M. Liwicki and H. Bunke. Combining on-line and off-line systems for handwriting recognition. In *Proc. ICDAR 2007, Vol. 1*, pp. 372–376, 2007. IEEE.

[4] R. Mayer and R. Neumayer. Multi-modal Analysis of Music: A large-scale Evaluation. In *Proc. WEMIS 2009*, pp. 30–35, 2009.

[5] C. McKay and I. Fujinaga. Automatic genre classification using large high-level musical feature sets. In *Proc. ISMIR 2004*, pp. 525–530, 2004.

[6] T. Pérez-García, J. M. Iñesta, and D. Rizo. metamidi: a tool for automatic metadata extraction from midi files. In *Proc. WEMIS 2009*, pp. 36–40, Oct. 2009.

[7] C. Pérez-Sancho, D. Rizo, and J. M. Iñesta. Genre classification using chords and stochastic language models. *Connection Sci.*, 21(2 & 3):145–159, May 2009.

[8] G. Rigoll and S. Müller. Statistical pattern recognition techniques for multimodal human computer interaction and multimedia information processing. In *Proc. of the Int. Workshop on Speech and Computer Information Processing, in Survey Paper*, pp. 60–69, 1999.

[9] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *Signal Proc. Mag., IEEE*, 23(2):133–141, March 2006.