An Evaluation of Symbolic Feature Sets and Their Combination for Music Genre Classification

Hanna C. B. Piccoli, Carlos N. Silla Jr.
Computer Music Technology Laboratory
Federal University of Technology of Paraná (UTFPR-CP)
Cornélio Procópio, Brazil
{piccolihanna,carlos.sillajr}@gmail.com

Pedro J. Ponce de León, Antonio Pertusa
Department of Software and Computing Systems
University of Alicante
Alicante, Spain
{pierre,pertusa}@dlsi.ua.es

Abstract—**The automatic music genre classification task is an active area of research in the field of Music Information Retrieval. In this paper we use two different symbolic feature sets for genre classification and combine them using an early fusion approach. Our results show that early fusion achieves better classification accuracy than using any of the individual feature sets. Furthermore, when compared with some of the state of the art approaches using the same experimental conditions, early fusion of symbolic features is ranked the second best method.**

I. Introduction

Automatic music genre classification is one of the most popular topics within the Music Information Retrieval (MIR) community [1]. One of its main motivations is to build systems that are able to automatically categorize huge amounts of music collections. Furthermore, there are studies that show that music genre is one of the most used terms by the users of MIR search engines [2], [3], [4].

One of the most important studies in this topic was done by [5] which extracted thirty features from the music signal in order to train an automatic music genre classification system. In the MIR domain, methods that extract features directly from the music signal are known as content-based audio approaches. For a recent survey on this subject, the reader is referred to [6].

Besides content-based audio algorithms, there are also content-based symbolic approaches using a symbolic representation of the song instead of the audio signal. Moreover, there are some works in the literature that use other types of information (or modalities) such as cultural data [7], lyrics [8] or social tags [9].

One of the current MIR trends is to combine features extracted from different modalities. Typically content-based audio features are combined with lyrics [10], symbolic features [11], context features [12], or symbolic, cultural and lyrics features [13].

However, there is little research on the subject of combining different types of features extracted from the same modality. One of the few exceptions is the work of [14], where the authors combine different types of content-based audio features. The combination of different types of audio features improved the music genre classification accuracy[14]. For this reason, in this study we evaluate whether it is possible to improve the music genre classification accuracy by combining two different types of content-based symbolic features, which to the best of the authors knowledge has not been done before.

In this study, we experimentally verify that using an early fusion approach with two different types of content-based symbolic features is better than using the standalone content-based symbolic feature sets. We also show that early fusion with two different types of content-based symbolic features is ranked the second best approach when compared against some of the state of the art methods.

This study is organized as follows: Sec. II presents an overview of the proposed methodology. Sec. III presents the experimental setup for the experiments. The computational results and discussion are presented in Sec. IV and finally, in Sec. V the conclusions and future research directions are described.

II. System Overview

In this section we present an overview of the configurations used by the our content-based symbolic music genre classification system.

In the first and second configuration, the system works as presented in Fig. 1–(a) and 1–(b). These configurations are used as baseline approaches, and they are composed of three main stages: (1) Audio to Midi Transcription; (2) Feature Extraction; (3) Music Genre Classification. Note that in these configurations the system extracts different types of content-based symbolic feature sets.

In the third configuration, the method works as presented in Fig. 1–(c). This configuration is composed of four stages: (1) Audio to Midi Transcription; (2) Feature Extraction; (3) Early Fusion combination of the different types of content-based symbolic features; and (4) Music Genre Classification.

As it can be seen in Fig. 1, the main difference between these configurations is that (a) and (b) extract only one type of content-based symbolic feature set whereas in (c) the two types of features sets are extracted and combined using an early

978-1-4799-0652-9/13 $31.00 © 2013 IEEE
DOI 10.1109/SMC.2013.327

IEEE computer society

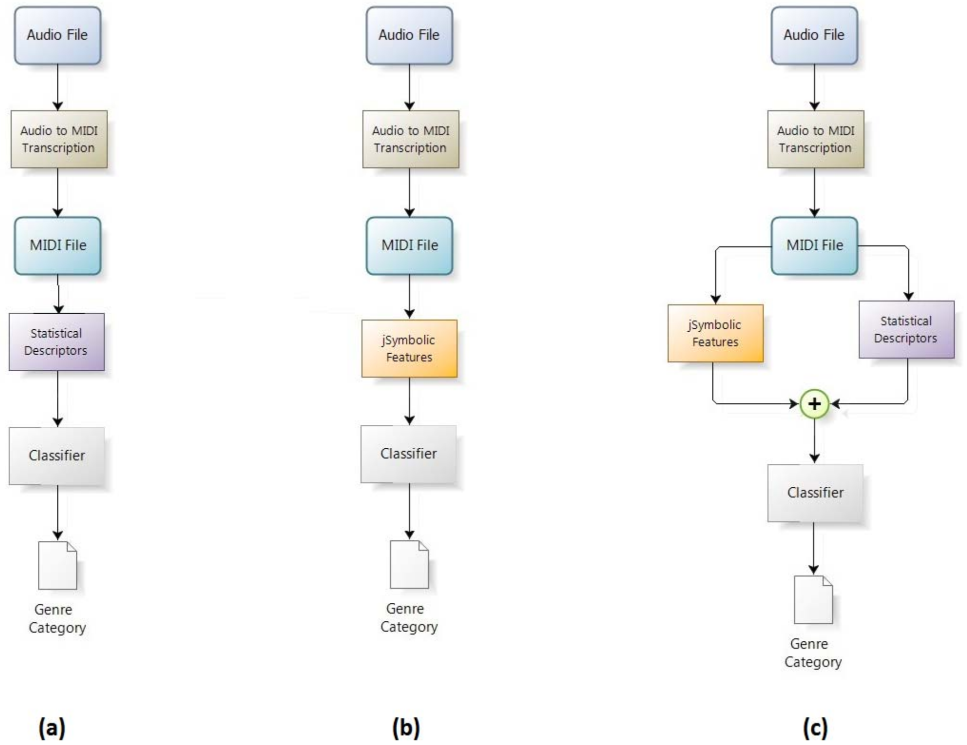

Fig. 1. Overview of the different configurations of the presented system.

fusion approach. In the following subsections we will briefly review the main components of the proposed content-based symbolic music genre classification system.

A. Audio to MIDI Transcription

In this study we have used a multiple fundamental frequency (f_0) method described in [15]. The output of this system is a single track MIDI file without any timbral information nor instrument separation, therefore a single track is yielded. The fundamental frequencies are estimated considering information about neighboring audio frames to get a smoothed temporal detection.

Although this system obtains state of the art results, music transcription is a challenging issue and the highest success rates with unknown timbres are around 60%[15]. However, besides these limitations, this scheme has been previously applied in [11] for genre classification, increasing the performance by introducing features on the symbolic level.

Given the nature of the transcriptor output, the working hypothesis is that it can help a music genre classification system based on symbolic features, by providing high level information, notably note pitches and durations. This transcription output, while not accurate, still contains enough information to meaningfully distinguish between genres.

B. Feature Sets

This section describes the two content-based symbolic feature sets used in this study.

1) Statistical Descriptors: This feature set is described in [16], [17], and it contains 44 statistical descriptors. These features are relative to overall statistics such as polyphony rate (proportion of sounding note periods with more than one note active simultaneously), average number of notes per beat, occupation rate (non-silence periods with respect to song length), song duration, and also to note pitches, pitch intervals, note durations, silence durations, inter-onset-intervals, non-diatonic notes, number of distinct pitch intervals, pitch interval mode, and estimation of the number of syncopations in the song.

2) jSymbolic Features: This symbolic feature set was obtained using the jSymbolic[1] framework [18]. This framework is a GUI-based java application for extracting features from MIDI files. It is bundled with several types of high-level music abstractions and its features fall into the broad categories of instrumentation, texture, rhythm, dynamics, pitch statistics and melody. In this study we have used this framework to extract a content-based symbolic feature set with 1022 features.

[1] Available at: http://jmir.sourceforge.net/jSymbolic.html

C. Early Fusion

In machine learning, a common approach when using multiple feature sets is to fuse all them in a single, multimodal feature space. This technique is known as early fusion and it is defined in [19] as a fusion scheme that integrates unimodal features before learning concepts. By using early fusion with the statistical descriptors and the jSymbolic features, a symbolic feature set with 1066 features is obtained.

The main advantages of early fusion are that correlation between features extracted from different modalities can be exploited to improve the classification performance, and only one (multimodal) model needs to be learned.

As mentioned earlier in the context of content-based symbolic music genre classification, the authors are unaware of any papers that may have used an early fusion approach with different types of content-based symbolic feature sets.

D. Classification

A support vector machine (SVM) classifier has been used. This classifier creates a maximum hyper-plane that divides two regions in the feature space. It is commonly used in two class problems, and for that reason it is needed to use some decomposing strategy to handle multi-class problems. In this work we have chosen pairwise classification as the decomposing scheme for a linear support vector machine trained with the sequential minimum optimization algorithm [20] using a polynomial kernel.

III. Experimental Details

This section presents the experimental settings of our work.

A. Dataset

The Latin Music Database (LMD) [21] has been used to perform the symbolic music genre classification experiments. The LMD contains 3227 MP3 music pieces from 10 different Latin genres (Axé, Bachata, Bolero, Forró, Gaúcha, Merengue, Pagode, Salsa, Sertaneja, Tango) originated from music pieces of 501 artists. In this database music genre assignment was manually annotated by a group of human experts, based on the human perception of how each music is danced. The genre labeling was performed by two professional teachers with over 10 years of experience in teaching ballroom Latin and Brazilian dances. It should be noted that the LMD is a benchmark database that has been used in the MIREX (Music Information REtrieval eXchange) [22].

B. Experimental Setup

The results reported in Sec. IV have been performed using exactly the same experimental procedure than in [23], [24], [25]. This is important in order to compare the results of our approach with some of the state of the art methods.

A three-fold cross-validation procedure with an artist filter [26] has been used. The reason for using three-fold instead of the ten-fold cross-validation procedure is because of the restrictions imposed by the artist filter. This filter makes that songs from a given artist must all be in the same fold during cross-validation. This is to avoid that instead of performing automatic music genre classification the system will learn to perform automatic artist identification. Due to the artist filter restriction the experiments were performed using 900 songs of the LMD. The songs were all equally distributed among the 10 music genres.

IV. Experiments

We are interested in answering the following questions: (a) Is it possible to improve the classification accuracy of content-based symbolic music genre classifier by using an early fusion approach with two different types of features? (b) How good is the performance of the early fusion of symbolic features when compared to previous works? All the experiments reported in this section were performed by using the experimental details previously presented.

A. Individual Feature Sets Vs. Early Fusion

We first evaluate whether it is better to use one of the individual symbolic feature sets or to use early fusion which combines both of them.

Tab. I presents the accuracy of each symbolic feature set employed in this study for each genre in the LMD, and the overall results. The accuracy is computed as the number of correctly classified instances divided by the total number of instances.

The analysis of the results in Tab. I reveals that the jSymbolic feature set is better than the statistical descriptors for classifying nine out of the ten Latin music genres in the LMD, with the exception of the music genre ”Merengue”.

The fourth column in Tab. I presents the results for the early fusion approach combining both the statistical and jSymbolic feature sets. The comparison of the early fusion feature set against both the individual feature sets shows that it improves the overall music genre classification accuracy.

As the number of samples is large enough, according to the central limit theorem, it can be assumed that the difference between two average success rates, p_1 and p_2, follows a normal distribution. The average success results have been obtained by a cross-validation setup, as stated above. The hypothesis test is $H_0 \equiv \widehat{p_1} - \widehat{p_2} \sim N(0, \sigma^2)$, with $\sigma^2 = p_1(1-p_1)/n + p_2(1-p_2)/m$, where $n = m = 900$ samples, that is, both success rates are not significantly different. The statistics $z = (\widehat{p_1} - \widehat{p_2})/\sigma$ is used to perform the significance test. The threshold value $z_{\alpha/2}$ for accepting the hypothesis with a level of significance $\alpha = 0.05$, is set to 1.96 for infinite degrees of freedom (given the large number of samples). If $|z| > z_{\alpha/2}$, the initial hypothesis is rejected at the α level. These are the testing conditions when comparing success rates throughout this paper.

Comparing average success rates for statistical and jSymbolic features, $|z| = 4.6 > z_{\alpha/2}$, hence the success rates are significantly different. However, comparing the jSymbolic and the early fusion setups gives $|z| = 1.43 < z_{\alpha/2}$, which is not a significant diference at the 5% significance level.

TABLE II
COMPARISON OF THE BEST RESULTS ACHIEVED IN THIS WORK AGAINST THE STATE OF THE ART REPORTED IN [25]

Music Genre	Early Fusion of Symbolic Features	Content-Based Audio Features [25]	Instance selection with Content-Based Audio Features [23]	LBP Features with Melscale zoning [25]	GLCM Features [24]	Instance Selection with GLCM Features [24]
Axé	67.76	57.78	61.11	83.33	73.33	76.67
Bachata	90.00	85.56	91.11	93.33	82.22	87.78
Bolero	72.23	63.33	72.22	91.11	64.44	83.33
Forró	52.23	38.89	17.76	82.22	65.56	52.22
Gaúcha	66.66	51.11	44.00	67.78	35.56	48.78
Merengue	84.46	78.89	78.78	95.56	80.00	87.78
Pagode	73.33	46.67	61.11	71.11	46.67	61.11
Salsa	88.90	57.78	40.00	84.44	42.22	50.00
Sertaneja	61.13	42.22	41.11	67.78	17.78	34.44
Tango	83.33	87.78	88.89	86.67	93.33	90.00
Overall	74.00	61.00	59.67	82.33	60.11	67.20

TABLE I
ACCURACY OF THE DIFFERENT FEATURE SETS IN THE LMD.

	Feature Set		
Genre	Statistical	jSymbolic	Early Fusion
Axé	51.10	63.33	67.76
Bachata	80.00	87.80	90.00
Bolero	68.90	66.66	72.23
Forró	34.43	43.36	52.23
Gaúcha	41.10	71.10	66.66
Merengue	84.46	77.76	84.46
Pagode	43.33	74.46	73.33
Salsa	74.43	85.53	88.90
Sertaneja	50.00	56.66	61.13
Tango	80.00	83.33	83.33
Overall	60.77	71.00	74.00

B. Early Fusion with Symbolic Features Vs. State of the Art

The performance of the early fusion approach with symbolic features has been compared to some of the state of the art methods. As mentioned in Sec. III, our experiments can be directly compared with previous methods using the same experimental conditions.

Tab. II presents the comparison of our early fusion method using content-based symbolic features with some of the previous works in the literature.

The second column of Tab. II presents the results for the early fusion approach combining two types of content-based symbolic features. These results are the same than those reported in Tab. I.

The third column of Tab. II presents the results using content-based audio features extracted with the Marsyas framework [5]. It should be noted that this framework uses different types of contend-based audio features, therefore it performs an early fusion of content-based audio features.

The fourth column of Tab. II shows the results using instance selection with content-based audio features [23]. This method uses only a portion of the available training data to achieve, at least, a similar performance as to when the whole training data is used. In its experiments, the Marsyas framework was used to extract content-based audio features.

The fifth, sixth and seventh columns of Tab. II show the results for a recent work that faces the problem from a different perspective. In [24], [25], the authors transform the audio signal into spectrograms and extract content-based features from the time-frequency images, performing music genre classification as an image classification problem. It should be noted that they also use time decomposition[27] for automatic music genre classification which is a method where the audio signal is segmented into a given number of pieces and the final classification is achieved by using late fusion. In their experiments they have used three segments, one for the beginning of the song, one for the middle, and one for the end. They have also used different rules for performing the late fusion approach. An example of a late fusion rule is the majority voting.

The difference between the results presented in the fifth, sixth and seventh columns of Tab. II are as follows: In the fifth column the authors [25] employ features extracted with Local Binary Patterns (LBP) and Melscale zoning. In the sixth column, the authors [24] employ features extracted with Gray Level Co-occurence Matrix (GLCM). In the seventh one, they employ the instance selection approach with Gray Level Co-occurence Matrix[24].

The analysis of the results in Tab. II shows some interesting conclussions. Our early fusion approach with content-based symbolic features obtains significantly better results than: using content-based audio features extracted with the Marysas framework [5] ($|z| = 5.95$); using instance selection ($|z| = 6.53$), which was a technique developed in [23] to build more reliable music genre classification systems; It also gets better accuracy than converting the music signal into spectograms and using Gray Level Co-occurrence Matrix (GLCM) feature descriptors [24] ($|z| = 6.34$) , even if they are used in combination with the instance selection technique ($|z| = 3.18$).

Our approach was only outperformed by the system proposed in [25] ($|z| = 4.30$), where a two-step late fusion is performed. First, a time decomposition approach is used, where the audio signal is divided into three short 30 seconds music segments from the beginning, middle and end of the song respectively. Then, for each segment they employ the Melscale zoning which consists of dividing the spectograms into a given number of pieces. After this procedure, they train one classifier for each "zone" and combine their results by employing different late fusion combination rules.

V. Conclusions

In this paper we have evaluated two different types of content-based symbolic features sets and their combination for music genre classification of latin music. The results show that using early fusion (i.e. concatenating two different types of content-based symbolic feature sets together) achieves better accuracy than using any of the two symbolic feature sets individually , though this difference it is not significant. This is probably due to the unbalanced size of the respective feature sets.

When evaluating our approach with the current state of the art (under the same experimental settings), our early fusion approach with content-based symbolic features achieves the second best classification accuracy overall, even when compared to more complex methods.

In this study we have only used one dataset, known as the Latin Music Database, and for this reason in future research we intend to evaluate our approach with more music genre classification datasets. We also plan to extend this study to investigate the use of content-based symbolic features for hierarchical music genre classification [28]. Another interesting research direction is to compare the performance of the early fusion method used in this work with the late fusion employed by several authors in previous studies [11], [14], [25], [29].

Acknowledgment

The authors from the Federal University of Technology of Paraná are supported by Fundação Araucária–a Brazilian research-support agency. The authors from the University of Alicante are supported by the Consolider Ingenio 2010 program (MIPRCV, CSD2007-00018), the Spanish Ministry project TIN2009-14247-C02-02 (DRIMS), and the Instituto Universitario de Investigacin en Informtica (IUII).

We would like to thank the authors of [25] for kindly sharing the information about the song artist and song title used in their three-fold cross-validation experiments.

References

[1] C. McKay and I. Fujinaga, "Musical genre classification: Is it worth pursuing and how can it be improved?" in *Proc. of the 7th Int. Conf. on Music Information Retrieval*, 2006, pp. 101–106.

[2] J. S. Downie and S. J. Cunningham, "Toward a theory of music information retrieval queries: System design implications," in *Proc. of the 3rd Int. Conf. on Music Information Retrieval*, 2002, pp. 299–300.

[3] J. H. Lee and J. S. Downie, "Survey of music information needs, uses, and seeking behaviours: preliminary findings," in *Proc. of the 5th Int. Conf. on Music Information Retrieval*, 2004, pp. 441–446.

[4] E. Pampalk, A. Rauber, and D. Merkl, "Content–based organization and visualization of music archives," in *Proc. of the 10th ACM Multimedia Conf.*, 2002, pp. 570–579.

[5] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[6] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303–319, 2011.

[7] B. Whitman and P. Smaragdis, "Combining musical and cultural features for intelligent style detection," in *Proc. of the 3rd Int. Conf. on Music Information Retrieval*, 2002, pp. 47–52.

[8] R. Mayer, R. Neumayer, and A. Rauber, "Rhyme and style features for musical genre classification by song lyrics," in *In Proc. of the 9th Int. Conf. on Music Information Retrieval*, 2008.

[9] L. Chen, P. Wright, and W. Nejdl, "Improving music genre classification using collaborative tagging data," in *Proc. of the 2nd ACM Int. Conf. on Web Search and Data Mining*, 2009, pp. 84–93.

[10] R. Neumayer and A. Rauber, "Integration of text and audio features for genre classification in music information retrieval." in *Proc. of the 29th European conference on Information Retrieval*, 2007, pp. 724–727.

[11] T. Lidy, R. Mayer, A. Rauber, P. J. P. de Leon, A. Pertusa, and J. M. Inesta, "A cartesian ensemble of feature subspace classifiers for music categorization," in *Proc. of the 11th Int. Conf. on Music Information Retrieval*, 2010, pp. 279–284.

[12] J. Aucouturier, F. Pachet, P. Roy, and A. Beurivé, "Signal + context = better classification," in *Proc. of the 8th Int. Conf. on Music Information Retrieval*, 2007, pp. 425–430.

[13] C. McKay, J. A. Burgoyne, J. Hockman, J. B. L. Smith, G. Vigliensoni, and I. Fujinaga, "Evaluating the genre classification performance of lyrical features relative to audio, symbolic and cultural features," in *Proc. of the 11th Int. Conf. on Music Information Retrieval*, 2010, pp. 213–218.

[14] C. N. Silla Jr., A. L. Koerich, and C. A. A. Kaestner, "Improving automatic music genre classification with hybrid content-based feature vectors," in *Proc. of the 25th ACM Symposium on Applied Computing*, 2010, pp. 1702–1707.

[15] A. Pertusa and J. M. Inesta, "Efficient methods for joint estimation of multiple fundamental frequencies in music signal," *EURASIP Journal on Advances in Signal Processing*, vol. 27, pp. 1–13, 2012.

[16] D. Rizo, P. J. P. de Leon, C. Perez-Sancho, A. Pertusa, and J. M. Inesta, "A pattern recognition approach for melody track selection in midi files," in *Proc. ISMIR, Victoria, Canada, 2006.*, 2006.

[17] P. J. P. de Leon and J. M. Inesta, "A pattern recognition approach for music style identification using shallow statistical descriptors," *IEEE Trans. on Systems Man and Cybernetics C*, vol. 37, no. 2, pp. 248–257, 2007.

[18] C. McKay and I. Fujinaga, "jSymbolic: A feature extractor for MIDI files," in *Proc. of the Int. Computer Music Conference*, 2006, pp. 302–305.

[19] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. of the 13th annual ACM Int. Conf. on Multimedia,*, 2005, pp. 399–402.

[20] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to platts smo algorithm for svm classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637–649, 2001.

[21] C. N. Silla Jr., A. L. Koerich, and C. A. A. Kaestner, "The latin music database," in *Proc. of the 9th Int. Conf. on Music Information Retrieval*, 2008, pp. 451–456.

[22] J. S. Downie, "The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[23] M. Lopes, F. Gouyon, A. L. Koerich, and L. E. S. Oliveira, "Selection of training instances for music genre classification," in *Proc. of the Int. Conf. on Pattern Recognition*, 2010, pp. 4569–4572.

[24] Y. M. G. Costa, L. S. Oliveira, A. L. Koerich, and F. Gouyon, "Music genre recognition using spectrograms," in *Proc. of the 18th Int. Conf. on Systems, Signals and Image Processing*, 2011, pp. 151–154.

[25] Y. M. G. Costa, L. E. S. Oliveira, A. L. Koerich, F. Gouyon, and J. G. Martins, "Music genre classification using lbp textural features," *Signal Processing*, vol. 92, no. 11, pp. 2723–2737, April 2012.

[26] A. Flexer, "A closer look on artists filters for musical genre classification," in *Proc. of the Int. Conf. on Music Information Retrieval*, 2007, pp. 341–344.

[27] C. N. Silla Jr., A. L. Koerich, and C. A. A. Kaestner, "A machine learning approach to automatic music genre classification," *Journal of the Brazilian Computer Society*, vol. 14, no. 3, pp. 7–18, 2008.

[28] C. N. Silla Jr. and A. A. Freitas, "Novel top-down approaches for hierarchical classification and their application to automatic music genre classification," in *Proc. of the IEEE Int. Conf. on Systems, Man, and Cybernetics*, 2009, pp. 3599–3604.

[29] C. McKay and I. Fujinaga, "Combining features extracted from audio, symbolic and cultural sources," in *Proc. of the 9th Int. Conf. on Music Information Retrieval*, 2008, pp. 597–602.