Brave New Task: MusiClef Multimodal Music Tagging

Cynthia C. S. Liem
Delft University of Technology
Delft, The Nederlands
c.c.s.liem@tudelft.nl

Nicola Orio
University of Padova
Padova, Italy
nicola.orio@unipd.it

Geoffroy Peeters
UMR STMS IRCAM-CNRS
Paris, France
geoffroy.peeters@ircam.fr

Markus Schedl
Johannes Kepler University
Linz, Austria
markus.schedl@jku.at

ABSTRACT

In this paper, we give an overview of the MusiClef Brave New Task at MediaEval 2012. After introducing the background of the task and related initiatives in the music information retrieval domain, we give an overview of the actual task and baseline results obtained for a reference implementation by the task organizers.

1. INTRODUCTION

MusiClef started as a lab at CLEF 2011 [5], focusing on a concrete real-world use case: automatic tagging of music items for production purposes. The CLEF lab led to a multimodal data collection for which the ground truth labeling was performed by use case stakeholders. Subsequently, for MediaEval 2012 a multimodal music tagging Brave New Task was proposed based on this collection. Unique properties about this task are that the released audio features are computed using a publicly available implementation, that other audio features can be computed on demand, and that the task has explicit interest in multimodal approaches. In this paper, we give an overview of the task; for a more elaborate introduction, the interested reader is referred to [4].

2. RELATED INITIATIVES

In the music information retrieval (MIR) domain, several benchmarking initiatives exist already, the first and most established initiative being the *Music Information Retrieval Evaluation eXchange* (MIREX) [2]. Due to copyright restrictions, access to datasets for MIREX frequently remains restricted to the MIREX organization, causing difficulties in replicating results of previous campaigns, and a dependence of obtained results on not just an algorithmic implementation, but also characteristics of locally acquired individual training sets.

Another recent relevant initiative to overcome music dataset sharing limitations is the *Million Song Dataset* (MSD). With the MSD, researchers can access a number of features from a very large song collection [1]. However, the feature set is fixed and the used feature extraction algorithms are not fully public, limiting possibilities to carry out further research on content description techniques. In 2012, the MSD launched a challenge on music recommendation for which multimodal and additional information sources may be used.

Finally, we mention the *Quaero* program on promoting research and innovation in multimedia indexing technologies. Every year, the Quaero-Eval initiative evaluates various audio and music processing tasks, the methology being inspired by both NIST and MIREX evaluations. Evaluation tasks, corpora and performance measures are defined upon common agreement. Algorithms are evaluated by an independent party. During an adjudication period, participants can check and discuss their results.

3. MULTIMODAL MUSIC TAGGING TASK

Music auto-tagging is the process of automatically assigning labels to music items. Such labels, or tags, can then be used for manifold music retrieval tasks, such as search, browsing and visualization. Most existing auto-tagging approaches for music take into account only one modality, either content-based [7] or text-based features [6].

The goal of the multimodal music tagging task is to exploit both *automatically extracted information about the content* and *user-generated data about the context* to carry out a tagging task: given the audio content of a song, a set of social tags associated to that song, and a set of web pages associated to the artist that performed the song, participants have to highlight the tags that best describe the song. It is not mandatory, although encouraged, to use all the sources of information.

The released MusiClef test collection consists of five parts:

1. songs: 1355 popular songs, recorded by 218 artists, split in a training set of 975 songs and a test set of 380 songs.
2. audio features: MFCC features computed with the MIRToolbox. Other features could be computed on demand.
3. user tags: social web tags for all songs, crawled from the `last.fm` music service.
4. web pages: crawled web pages on artists and albums in multiple languages. Together with the crawled web pages, standard ($tf \cdot idf$) values were provided.
5. ground truth: tag annotations made by stakeholder music professionals. A vocabulary of 94 tags was used.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

strategy	accuracy	recall	precision	specificity	F-measure
audio	0.894	0.148	0.127	0.939	0.126
user tags	0.898	0.061	0.039	0.942	0.0370
web pages	0.897	0.050	0.007	0.954	0.0110
majority	0.880	0.123	0.086	0.922	0.0860
union	0.824	0.240	0.115	0.845	0.1340

Table 1: Evaluation results for different data approaches, considered over the full dataset.

4. REFERENCE IMPLEMENTATION

In order to assist the participants in interacting with the provided data and to establish a simple baseline, a reference implementation in Matlab was made by the organizers. This implementation adopted the straightforward approach of training individual Gaussian Mixture Models for the audio features, user tags and web page data, and applied classification through a 1-nearest neighbor approach, based on symmetrized Kullback-Leibler divergence.

5. EVALUATION PROCEDURE

We adopted evaluation measures which are common in the information retrieval domain: accuracy, recall, precision, specificity and F-measure.

The tag vocabulary is of a large diversity. For example, it contains tags such as 'countryside', 'hopeful', 'reggae' and 'travel', which all express different musical aspects and use cases. We conjectured that these different aspects would maybe need different types of approaches in terms of modalities, and therefore made a functional categorization of the tags to allow a deeper analysis of this, a.o. including categories related to affect, genre, sound quality, but also specific occasions or places for which the song would be appropriate. More details on the categorization is given in [4].

6. BASELINE RESULTS

Employing our reference implementation, we ran evaluations for several strategies, obtaining results for five cases: (1) consideration of *audio* features only, (2) consideration of *user tag* features only, (3) consideration of *web page* features only, (4) *Majority* vote, considering all three data resources, and only keeping tags indicated by at least two of the resources, and (5) *Union*: taking the union of the tags obtained for each of the three data resources.

Evaluation results for these different scenarios considering the full dataset are shown in Table 1. Results on the F-measure for these scenarios, considering different tag categories, are shown in Figure 1. From these results, it can be seen that for our simple approach, the textual resources perform strongly inferior to the audio resources in case of a single-resource approach. Given that in terms of fusion strategies, the union performs generally better than the majority vote, the different resources appear to yield different tags. When considering the union versus an audio-only approach, it is seen that the audio-only approach only performs superior on explicitly audio-related tags (categories sound: temporal and sound: timbral). The physical situation tag category appears to benefit the most from the addition of textual data. However, additional results from other strategies would be needed before full conclusions on this can be drawn.

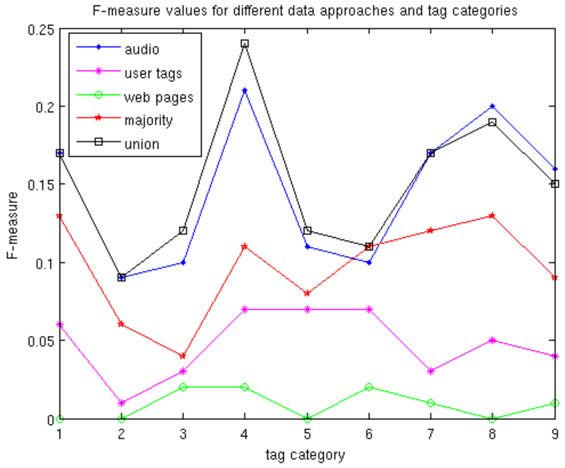

Figure 1: F-measure values for different data approaches (indicated in the legend), considered per tag category. The correspondence between horizontal axis indices and categories is: (1) activity/energy, (2) affective state, (3) atmosphere, (4) other, (5) situation: occasion, (6) situation: physical, (7) sociocultural: genre, (8) sound: temporal, (9) sound: timbral.

Acknowledgments

The authors would like to thank David Rizo, of the University of Alicante, for his support in starting the MusiCLEF initiative. MusiClef has been partially supported by the PROMISE Network of Excellence, co-funded by EU-FP7 (grant no. 258191), by the Quaero Program funded by Oseo French agency and by the MIReS project funded by EU-FP7-ICT-2011.1.5-287711, and by the Austrian Science Funds (FWF): P22856-N23. The work of Cynthia Liem is supported in part by the Google European Doctoral Fellowship in Multimedia.

7. REFERENCES

[1] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The Million Song Dataset. In *Proc. ISMIR*, 2011.

[2] J. S. Downie, K. West, A. F. Ehmann, and E. Vincent. The 2005 Music Information Retrieval Evaluation Exchange: Preliminary Overview. In *Proc. ISMIR*, 2005.

[3] O. Lartillot and P. Toiviainen. A Matlab Toolbox for Musical Feature Extraction from Audio. In *Proc. DAFx*, 2007.

[4] N. Orio, C. C. S. Liem, G. Peeters, and M. Schedl. MusiClef: Multimodal Music Tagging Task. In *Proc. CLEF*, 2012.

[5] N. Orio, D. Rizo, R. Miotto, N. Montecchio, M. Schedl, and O. Lartillot. MusiCLEF: A Benchmark Activity in Multimodal Music Information Retrieval. In *Proc. ISMIR*, 2011.

[6] M. Schedl and T. Pohle. Enlightening the Sun: A User Interface to Explore Music Artists via Multimedia Content. *MTAP*, 49(1), August 2010.

[7] K. Seyerlehner, M. Schedl, P. Knees, and R. Sonnleitner. A Refined Block-Level Feature Set for Classification, Similarity and Tag Prediction. In *Extended Abstract to MIREX*, 2011.