MusiClef: Multimodal Music Tagging Task

Nicola Orio, Cynthia C. S. Liem, Geoffroy Peeters, and Markus Schedl

University of Padua, Italy; Delft University of Technology, The Netherlands; UMR STMS IRCAM-CNRS, Paris, France; Johannes Kepler University, Linz, Austria
musiclef@dei.unipd.it

Abstract. MusiClef is a multimodal music benchmarking initiative that will be running a MediaEval 2012 Brave New Task on Multimodal Music Tagging. This paper describes the setup of this task, showing how it complements existing benchmarking initiatives and fosters less explored methodological directions in Music Information Retrieval. MusiClef deals with a concrete use case, encourages multimodal approaches based on these, and strives for transparency of results as much as possible. Transparency is encouraged at several levels and stages, from the feature extraction procedure up to the evaluation phase, in which a dedicated categorization of ground truth tags will be used to deepen the understanding of the relation between the proposed approaches and experimental results.

1 Introduction

MusiClef is a benchmarking activity that will run as a Brave New Task in MediaEval 2012. Brave New Tasks are a new category of MediaEval tasks, meant to pilot promising and new, but potentially risky tasks. After creating a test collection as a lab at CLEF 2011 [10], the collection will now be used for a multimodal Music Information Retrieval (MIR) benchmarking activity in MusiClef 2012.

MusiClef is built around a *concrete real-world use case* centered around music production. Stakeholders from this domain were involved in the original ground truth labeling, and will remain involved at the evaluation phase.

Although copyright restrictions prevent original music audio to be shared, MusiClef aims at allowing *replication of the results* by distributing both content features and the algorithms used to extract them. An initial set of features, based on open source implementations of music processing techniques, is provided to participants. Additionally, it will be possible for participants to propose alternative features that will then be computed on-demand.

Finally, MusiClef promotes *multimodal approaches* on the music objects. As has been suggested before in the community [7], approaches going beyond audio signal content may be necessary to properly address and solve real-world use cases. Thus, besides audio features, related information in the form of social tags and web pages will be provided, and participants are encouraged to include other modalities and sources of additional information in their approaches.

2 Related Initiatives

The Music Information Retrieval Evaluation eXchange: The need for shared evaluation practices has been clear in the MIR community since 2004, when a first campaign on audio feature extraction was organized by Pompeu Fabra University at the ISMIR conference. From the year after, a very important evaluation campaign for this research was started by the University of Illinois: the Music Information Retrieval Evaluation eXchange (MIREX) [4]. Due to copyright restrictions, the organizers of the MIREX can only distribute publicly available test collections. For the rest, participants must locally experiment on their own test collections, after which they submit their software to be run on the evaluation set by the organizers. This approach has two drawbacks, which have already been debated by the MIR research community: the results of previous campaigns cannot be easily replicated and the performances depend on the individual training sets and not only on the submitted algorithms.

The Million Song Dataset Challenge: A 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 [3]. 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[1] on music recommendation for which, similarly to MusiClef, multimodal and additional information sources may be used. However, despite similarities between MusiClef and the MSD challenge and the much larger corpus size of the MSD, MusiClef still validly offers a complementary alternative. With the professional use case from which the MusiClef corpus was built, manual labels attached to MusiClef items will be much cleaner than those of the MSD corpus, and more relevant to the dedicated practical use case. Furthermore, as indicated above, while not being able to publicly share audio data, MusiClef allows audio feature (re)computation on demand, allowing advancement on content description techniques too.

Quaero-Eval: Quaero is a program promoting research and industrial innovation on technologies for automatic analysis and classification of multimedia and multilingual documents gathering around 30 French and German public and private research organizations. Evaluation plays an important role in the program. In particular, Quaero-Eval focuses on audio and music processing, inspired by NIST and MIREX evaluations. Tasks to be run are defined upon common agreement, as are the annotated corpus to be used, the evaluation measures and the way the results will be published. A Mercurial repository allows participants to share and test the implementation of the evaluation framework and to access the training part of the annotated corpus. Submitted algorithms are run on the test sets using evaluation frameworks by an independent body that does not participate in the evaluation. Results are then communicated to the participants. After

[1] `http://www.kaggle.com/c/msdchallenge`

the evaluation has been run, the test sets are made public and an adjudication period starts in which participants can check in detail their results and comment on the annotations of the test sets. For each task, a report detailing the results is then written. A post-evaluation meeting allows participants to discuss in detail the results obtained during the campaign. The test set used for a given year becomes the training set of the following year. For comparison purposes, evaluation can also be performed on the test-sets of the previous years.

MediaEval MediaEval[2] is a relatively young, but rapidly growing benchmarking initiative that focuses on human and social aspects of multimedia. Originally established in 2008 as VideoCLEF, a track within CLEF focusing on the analysis of and access to multilingual multimedia content, it became an independent benchmarking initiative in 2010, adopting the name MediaEval and expanding the number of tasks. MediaEval strives to emphasize the *multi* in multimedia, including the use of speech, audio, tags, users, context as well as visual content. Because of this emphasis, MediaEval attracts a diverse group of researchers, both from industry and academia, with a large range of perspectives on multimedia research. MediaEval works by exploiting this diversity to drive innovation in task design and data collection development [5]. The main risk of MusiClef in the MediaEval context is that music currently is not commonly seen as multimedia data. However, we are strongly convinced that open challenges in music and multimedia research are very much alike [7], and thus intend to attract a multidisciplinary audience to the MusiClef benchmarking task.

3 Multimodal Music Tagging Task

Music auto-tagging is the process of automatically assigning semantic labels to music items (e.g., songs or artists). Such labels, or tags, can then be used for manifold music retrieval tasks, for instance, semantic text-based music search and faceted browsing of music collections, as well as for creating multimodal visualizations of music repositories. Typically, a machine learning approach, a *supervised learner*, is employed on a training data set to associate feature representations of music pieces with semantic tags. After training is finished, the classifier is used to predict labels to previously unseen music items. Most existing auto-tagging approaches for music take into account only one modality. Typically, content-based features extracted from the audio signal are used, for instance in [12,13]. Relying only on contextual, text-based features, a dictionary of music terms is used to index web pages and in turn assign tags to music artists in [11]. Mandel et al. [9] learn tag language models over different sets of vocabularies. With MusiClef, we aim at fostering multimodal approaches.

Task: 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

[2] `http://www.multimediaeval.org`

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 task is based on a real application scenario: songs of a commercial music library need to be categorized according to their possible usage in TV and radio broadcasts or web streaming (commercials, soundtracks, jingles). When this task is carried out manually, as it is still done by many companies, it is typical to exploit both audio content and contextual information.

Test Collection: The test collection consists of five parts:

– *Songs:* Because of the focus on multimodality, all the different sources of information should give a comparable contribution to the tagging task. Hence, one of the requirements for the test collection was to select well-known songs by popular artists. This way, we can expect that enough social tags are available for each song and enough web pages are available for each artist. We collected the songs starting from the "Rolling Stone 500 Greatest Songs of All Time", which lists songs that have been recorded by a total of 218 different artists. The initial list of 500 songs was increased by adding at most 8 songs for each artist, obtaining a final list of 1355 songs.[3]

– *Audio features:* For copyright reasons, content descriptors are made available through the distribution of audio features computed using the publicly available MIRtoolbox [6]. Participants may also request to use specialized features, and can submit their own feature extraction algorithms for this.

– *User tags:* The web service made available by `last.fm` has been used to automatically gather the user tags associated to each song. Tags are in the form of a simple list of terms.

– *Web crawling:* To offer another kind of contextual data, we performed web crawls using a major search engine to retrieve the URLs of the top-ranked pages for queries including artist and album names. Fetching the web pages corresponding to these URLs, we are able to provide music-related sets of web pages in different languages.

– *Ground truth:* Each song in the dataset has been manually annotated by music professionals, who routinely add textual descriptors to commercial music libraries. The vocabulary of tags was initially composed of 355 tags: 167 for genre and 288 for mood. Manual tagging was carried out through a web interface, from which it was possible to listen to the complete songs and select the associated tags through a number of checkboxes, divided in genre and mood. Annotators were required to provide at least one tag for genre and five tags for mood. From the initial set, we kept only the tags that have been assigned to at least 10 songs, obtaining a final list of 94 tags.

[3] For this campaign we purposely excluded live versions and covers, because the former can have a variable audio quality and the latter can give inconsistencies between tags related to the performer and web pages related to the composer.

4 Evaluation Procedures

4.1 Applying a Deeper Ground Truth Tag Categorization

It has been acknowledged that the types of tags that users add to music can fall into different categories, which do not relate to audio signal content in equal ways [1,2,9]. While a social tag describing a featured instrument ('guitar') can be inferred from the signal, this will be much harder for a personal tag ('seen live'). This is also seen in the ground truth tagging vocabulary of MusiClef. Tags like 'travel', 'club', and 'ballroom' have strong contextual non-audio connotations.

Other evaluation initiatives did not explicitly consider in depth yet the existence of multiple tag categories. MusiClef will do this, aiming to advance transparency and deeper insight into how different categories of tags may imply different feature choices and tagging approaches. Based on the final ground truth tag set, we propose a categorization more specific than 'genre' and 'mood' for MusiClef, partially inspired by musicological theories on film music functions [8], and touching upon different music aspects and potential use cases:

1. *situation*, time and space aspects of the music:
   (a) *physical situation*: concrete physical environments (e.g. 'city', 'night').
   (b) *occasion*: implications of time and space, typically connected to social events (e.g. 'holiday', 'glamour').
2. *sociocultural genre*, belonging to a certain *style*, with dedicated social communities identifying with them (e.g. 'new wave', 'r&b', 'punk').
3. *affective*, mood-related aspects:
   (a) *activity*: the amount of perceived music activity, without implying strong positive or negative affective qualities (e.g. 'fast', 'mellow', 'lazy').
   (b) *affective state*: affective qualities that can only be connected and attributed to living beings (e.g. 'aggressive', 'hopeful').
   (c) *atmosphere*: affective qualities that can be connected to environments (e.g. 'chaotic', 'intimate').
4. *sound qualities*, aspects that can clearly be connected to audio signal content:
   (a) *timbral aspects* (e.g. 'acoustic', 'bright').
   (b) *temporal aspects* (e.g. 'beat', 'groove').
5. *other*, for tags not in the above categories (e.g. 'catchy', 'evocative').

Tags may fall into multiple categories. A first categorization for the ground truth tags was made by the MusiClef organizers. This will be further revised after discussion with the task participants. At the evaluation phase, evaluation measures will not just be computed for the full ground truth set, but also explicitly be considered in relation to the proposed categorization above.

4.2 Reference Implementation

Participants can take advantage of a reference implementation that will be made available by the organizers. This implementation has two main goals: serving as a starting point for setting up a development code framework, and creating a baseline for participants to compare the effectiveness of their approaches. The reference implementation will be based on state-of-the-art auto-tagging approaches, without optimizations to maintain transparency.

4.3 Evaluation Measures

For a specific set of tags (possibly grouped into sub-categories), performances of the systems will be measured using both threshold-based measures (binary relevance) and affinity measures. For the binary relevance (tag-based classification), accuracy, positive/negative example accuracy, precision, recall and f-measure will be considered as measures. The affinity measure will be based on the Area Under ROC Curve.

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.

References

1. Aucouturier, J.J.: Sounds Like Teen Spirit: Computational Insights into the Grounding of Everyday Musical Terms. In: Minett, J., Wang, W. (eds.) Language, Evolution and the Brain. Academia Sinica Press (2009)
2. Bertin-Mahieux, T., Eck, D., Mandel, M.: Automatic Tagging of Audio: The State-of-the-Art. In: Wang, W. (ed.) Machine Audition: Principles, Algorithms and Systems. IGI Publishing (2010)
3. Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P.: The Million Song Dataset. In: Proc. ISMIR (2011)
4. Downie, J.S., West, K., Ehmann, A.F., Vincent, E.: The 2005 Music Information Retrieval Evaluation Exchange: Preliminary Overview. In: Proc. ISMIR (2005)
5. Larson, M., Soleymani, M., Eskevich, M., Serdyukov, P., Ordelman, R., Jones, G.: The Community and the Crowd: Developing Large-scale Data Collections for Multimedia Benchmarking. IEEE MultiMedia (to appear, 2012)
6. Lartillot, O., Toiviainen, P.: A Matlab Toolbox for Musical Feature Extraction from Audio. In: Proc. DAFx (2007)
7. Liem, C.C.S., Müller, M., Eck, D., Tzanetakis, G., Hanjalic, A.: The Need for Music Information Retrieval with User-Centered and Multimodal Strategies. In: Proc. MIRUM (2011)
8. Lissa, Z.: Ästhetik der Filmmusik. Henschelverlag, Berlin, Germany (1965)
9. Mandel, M.I., Pascanu, R., Eck, D., Bengio, Y., Aiello, L.M., Schifanella, R., Menczer, F.: Contextual Tag Inference. ACM TOMCCAP 1(7S) (October 2008)
10. Orio, N., Rizo, D., Miotto, R., Montecchio, N., Schedl, M., Lartillot, O.: MusiCLEF: A Benchmark Activity in Multimodal Music Information Retrieval. In: Proc. ISMIR (2011)
11. Schedl, M., Pohle, T.: Enlightening the Sun: A User Interface to Explore Music Artists via Multimedia Content. MTAP 49(1) (August 2010)
12. Seyerlehner, K., Schedl, M., Knees, P., Sonnleitner, R.: A Refined Block-Level Feature Set for Classification, Similarity and Tag Prediction. In: Extended Abstract to MIREX (2011)
13. Sordo, M.: Semantic Annotation of Music Collections: A Computational Approach. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain (2012)