1000 Songs Database

Mohammad Soleymani[1], Michael N. Caro[2], Erik M. Schmidt[2], Cheng-Ya Sha[3] and Yi-Hsuan Yang[4]

[1]Department of Computer Science, University of Geneva, Switzerland
mohammad.soleymani@unige.ch
[2]ECE Department, Drexel University, USA
[3]National Taiwan University, Taiwan
[4]Academia Sinica, Taiwan

November 10, 2014

Abstract. This is a manual to help the users to use the database of 1000 songs for emotional analysis in music. The emotional annotations are collected with the goal of detecting the emotions that are expressed by the music and musicians form the content.

1. Introduction

1000 songs has been selected from Free Music Archive (FMA)[1]. The excerpts which were annotated are available in the same package song ids between 1 and 1000. We initially developed a dataset of 1,000 songs. However, a set of duplicates were later discovered and removed, which reduced the size of the dataset to 744 songs. The dataset is split between the development set (619 songs) and the evaluation set (125 songs). The extracted 45 seconds excerpts are all re-encoded to have the same sampling frequency, i.e, 44100Hz. Full songs are available at are also provided in the same package. The 45 seconds excerpts are extracted from random (uniformly distributed) starting point in a given song. The continuous annotations were collected at a sampling rate which varied by browsers and computer capabilities. Therefore, we resampled the annotations and generated the averaged annotations with 2Hz sampling rate. In addition to the average, we provide the standard deviation of the annotations so that you can have an idea about the margin of error. The continuous annotations are between -1 and +1 and excludes the first 15 seconds due to instability of the annotations at the start of the clips. To combine the annotations collected for the whole song, on nine points scale, we report the average and the standard deviation of the ratings ranging from [1, 9]. A detailed explanation of data collection methods as well as baseline results are provided in [4]. The submission results of the three teams who have participated in, Mediaeval 2013 Emotion in Music task, is available in the proceedings [2].

2. Data Collection

The songs were annotated by crowdworkers (annotators) on Amazon Mechanical Turk. Each song was annotated once for arousal and once for valence separately. The crowdworkers were asked to annotate the emotion music intends to induce and not the crowdworkers' own emotion. A video was made to instruct the workers on how to perform the annotations[2]. More than 300 crowdworkers on Amazon mechanical Turk were recruited after passing a qualification test to participate in our annotations. They received 0.40$ to annotate 3 songs on arousal and valence scales.

3. Data Format

All the songs received 10 annotations from which we provide the average and standard deviation of the scores given by Amazon MTurk workers. Annotations are made available in csv format. There are six csv files in this database, four containing average and standard deviation of arousal and valence continuous annotation for each song (second 15-45, `valence_cont_average.csv`,`valence_cont_std.csv`, `arousal_cont_average.csv`, `arousal_cont_std.csv`) with 2Hz sampling frequency.
"`whole_song_annotations.csv`" contains the average and standard deviation of the ratings given to the whole clip (static annotations). The last file, `songs_info` contains the metadata for all the songs, including, genre, title and musician.

[1]http://freemusicarchive.org/
[2]`https://www.youtube.com/watch?v=G-GhONd_Wag`

The development set and evaluation sets, utilized in Mediaeval 2013, are indicated in the `songs_info` file, "Mediaeval 2013 set" column.

Both the 45 seconds clips and full songs are provided in MPEG layer 3 (MP3) format.

4. features

A set of features, extracted by openSMILE[3] [1, 3] are provided.

The features are in WEKA's Arff format[4]. The compressed file, `default_features.zip`, contains the features for 500ms windows. The arff files contain a header describing the columns (features). After the @data field the data is contained as a comma separated table with rows representing feature vectors and columns features/attributes as described in the header. The first column always contains the clip name, the second column in the frames files contains the frame number. The last three columns contain the labels (arousal, valence, genre), which are filled with the ground truth values.

5. Baseline results

To evaluate the estimation models from content features the R^2 statistics and root-mean-square error (RMSE) are reported for static estimation and averaged correlation ($\bar{\rho}$) and RMSE are reported for dynamic estimation. Averaged correlation is a measure of the similarity of the trends and waveforms, whereas the RMSE provides an estimate of how far off the estimations were. The reported measures on dynamic annotated data are averaged for all the excerpts. Random level results are calculated by setting the target to the average score in the development set.

All the static and dynamic annotations and subsequently the predictions were scaled between $[-0.5, 0.5]$. A summary of the results, based on the systems proposed in the 2013 task [2] , is given in Table 1. Both in the static and dynamic task, the arousal estimations are far better than valence estimations. All the RMSE for dynamic estimations of valence and arousal for the three submissions are significantly lower (one-sided Wilcoxon test p-value<0.01) than the random level (averaged training targets) and than the Baseline. The simple Baseline system was able to perform better than random for arousal but fell short of performing better than random for valence estimation for both static and dynamic subtasks. However, their correlations vary by submissions and emotion dimensions. For the static subtask, all three submissions outperformed the provided Baseline. The dynamic subtask appeared to be more challenging and only TUM system [2] could consistently beat the baseline performance. The BLSMT-RNN takes advantage of temporal dependency, which is not supported by MLR, SVR or GPR that was used in the Baseline and Systems UoA and UU.

6. Terms of Use

The usage of the dataset is free of charge for non-profit research and we are by no means able to provide any support for its users. This database comes with no guaranty of correctness and we are certainly not liable to any damage it might cause. Any publication resulting from this dataset should acknowledge the authors by citing the paper published in CrowdMM workshop, ACM Multimedia 2013. The full reference is:

```
@inproceedings{1000SongforEmotioninMusic,
 author = {Soleymani, Mohammad and Caro, Micheal N. and Schmidt, Erik M. and Sha, Cheng-Ya
 and Yang, Yi-Hsuan},
 title = {1000 Songs for Emotional Analysis of Music},
 booktitle = {Proceedings of the 2Nd ACM International Workshop on Crowdsourcing for Multimedia},
 series = {CrowdMM '13},
 year = {2013},
 isbn = {978-1-4503-2396-3},
 location = {Barcelona, Spain},
 pages = {1--6},
 url = {http://doi.acm.org/10.1145/2506364.2506365},
 doi = {10.1145/2506364.2506365},
 publisher = {ACM},
```

[3] http://opensmile.sourceforge.net/

[4] http://www.cs.waikato.ac.nz/~ml/weka/arff.html

```
 address = {New York, NY, USA},
}
```

If you use the features provided in this dataset please acknowledge the contribution of its developers by citing the the following reference:

```
@inproceedings{openSMILE,
 author = {Eyben, Florian and Weninger, Felix and Gross, Florian and Schuller, Bj\"{o}rn},
 title = {Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor},
 booktitle = {Proceedings of the 21st ACM International Conference on Multimedia},
 series = {MM '13},
 year = {2013},
 isbn = {978-1-4503-2404-5},
 location = {Barcelona, Spain},
 pages = {835--838},
 url = {http://doi.acm.org/10.1145/2502081.2502224},
 doi = {10.1145/2502081.2502224},
 publisher = {ACM},
 address = {New York, NY, USA}
}
```

7. Acknowledgement

The recording of this dataset was not possible without the financial support from European Research Area under the FP7 Marie Curie Intra-European Fellowship: Emotional continuous tagging using spontaneous behavior (EmoTag).

Table 1. To evaluate the estimation models from content features the R^2 statistics and root-mean-square error (RMSE) are reported for static estimation and averaged correlation ($\bar{\rho}$) and RMSE for dynamic estimation. $\bar{\tau}$ is the normalized ranking distance of the retrieved songs using emotions. For RMSE and $\bar{\tau}$, smaller is better and for R^2 and $\bar{\rho}$ larger is better, where RMSE,$R^2, \bar{\tau} \in [0, 1], \bar{\rho} \in [-1, 1]$. Acronyms: RND: random level, BSL: Baseline,TUM: TU Munich, UoA: University of Aizu, UU: Utrecht University

(a) Static

Run	**Arousal**		**Valence**		**Ranking Dist.**
	RMSE	R^2	RMSE	R^2	$\bar{\tau}$
RND	0.16	0	0.15	0	0.43
BSL	0.12	0.48	0.15	0	0.40
TUM	0.10	0.59	0.11	0.42	0.34
UoA	0.10	0.63	0.12	0.35	0.35
UU	0.10	0.59	0.12	0.31	0.35

(b) Dynamic

Run	**Arousal**		**Valence**	
	RMSE	$\bar{\rho}$	RMSE	$\bar{\rho}$
RND	0.25 ± 0.13	0.10 ± 0.33	0.23 ± 0.11	0.05 ± 0.31
BSL	0.25 ± 0.11	0.16 ± 0.36	0.23 ± 0.10	0.06 ± 0.30
TUM	0.08 ± 0.05	0.31 ± 0.37	0.08 ± 0.04	0.19 ± 0.43
UoA	0.10 ± 0.05	0.11 ± 0.36	0.09 ± 0.05	0.06 ± 0.28
UU	0.10 ± 0.06	0.14 ± 0.28	0.12 ± 0.07	-0.01 ± 0.27

References

1. Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, *Recent developments in opensmile, the munich open-source multimedia feature extractor*, Proceedings of the 21st ACM International Conference on Multimedia (New York, NY, USA), MM '13, ACM, 2013, pp. 835–838.
2. Martha Larson, Xavier Anguera, Timo Reuter, Gareth J.F. Jones, Bogdan Ionescu, Markus Schedl, Tomas Piatrik, Claudia Hauff, and Mohammad Soleymani (eds.), *Working notes proceedings of the mediaeval 2013 workshop. Barcelona, Spain*, CEUR Workshop Proceedings, 2013.
3. Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al., *The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism*, Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, 2013.
4. Mohammad Soleymani, Micheal N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang, *1000 songs for emotional analysis of music*, Proceedings of the 2Nd ACM International Workshop on Crowdsourcing for Multimedia (New York, NY, USA), CrowdMM '13, ACM, 2013, pp. 1–6.