MUSIC EMOTION RANKING

Yi-Hsuan Yang and Homer H. Chen

National Taiwan University

ABSTRACT

Content-based retrieval has emerged as a promising approach to information access. In this paper, we propose an approach to music emotion ranking. Specifically, we rank music in terms of arousal and valence and represent each song as a point in the 2D emotion space. Novel ranking-based methods for annotation, learning, and evaluation of music emotion recognition are developed and tested on a moderately large-scale database composed of 1240 pop songs. Results are provided to show the feasibility of the proposed approach.

Index Terms— Music emotion recognition, rating, ranking

1. INTRODUCTION

Due to the explosive growth of music recordings, effective means for music retrieval and management is needed in the digital era. Classification and retrieval of music by emotion [1]–[5] has recently received increasing attention, because it is content-centric and functionally powerful.

A typical approach to music emotion recognition (MER) divides emotions into classes (e.g., happy, angry, sad and relaxing) and applies machine learning techniques to train a classifier. Each song is assigned a class label chosen among a handful of classes to represent the overall emotion embedded in the song. This categorical approach, however, faces the granularity issue that the precision needed for effective access of music in practice is higher. Using a finer granularity for emotion description does not necessarily address the issue since language is ambiguous and the description for the same emotion varies from person to person.

An alternative [6] is to view emotion from a continuous perspective and represent them in a 2D emotion space (2DES) in terms of arousal (how exciting/calming) and valence (how positive/negative), the two basic emotional dimensions found to be most important and universal [7]. Associated with the arousal and valence values (AV values), each song is represented as a point in the 2DES, and a user can easily retrieve music pieces of certain desired emotions by specifying a point or drawing a trajectory [5], as shown in Fig. 1, thereby alleviating the granularity and ambiguity issues of the categorical approach.

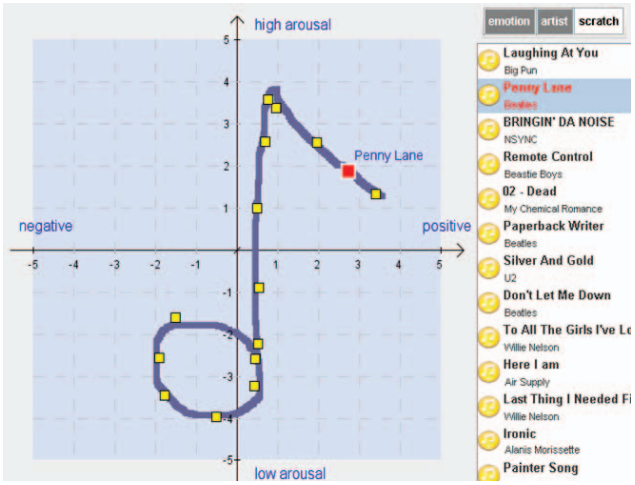

Fig. 1. Associated with the arousal and valence values, each song is represented as a point in the 2D emotion space, where a user can specify points or draw trajectories to retrieve music [5].

However, up to now, research in automatic prediction of the AV values is still at its early stage; many important issues are unaddressed. The first issue is related to the annotation of emotion. The *rating* measure (either the standard ordinal rating scale or the graphic rating scale) commonly adopted in the literature [3]–[8] may not be an appropriate choice due to the heavy cognitive load for human to directly express a continuum of emotions. In addition, it cannot ensure that the scale is consistent between and within subjects [9]. A score of 80 (out of 100) can mean fairly differently for two people.

The second issue is related to the learning of an MER system. Existing automatic approaches such as those in [3] and [4] use the mean square error (MSE) between predicted AV values and the ground truth as the objective function to train a model. However, due to the symmetric property of 2DES, such a model tends to give conservative estimate of AV values (distributed around the origin) [4], reducing the semantic coverage of the 2DES. More importantly, when representing songs in the 2DES, it is arguable whether human perceptually evaluates the accuracy with respect to the *absolute* position of each song or the position *relative* to one another. MSE cannot measure the latter one.

[1] This work is supported by a grant from the National Science Council of Taiwan under the contracts NSC 97-2221-E-002-111-MY3.

In light of the above observations, we propose a new approach that formulates MER as a *ranking* problem and trains a model to automatically rank music in terms of arousal or valence. We argue, and verify through subjective evaluation, that it is much easier for human to comprehend a continuum of emotions in a *comparative* way. For example, it is easier for us to tell which song is more exciting than to assign an exact arousal value to each song. The ordering of music pieces is aligned to the arousal or valence axis to generate the continuous representation over the 2DES.

New methodologies are needed to develop a ranking-based MER system that covers the annotation, learning, and evaluation processes. We develop a number of novel algorithms, incorporate them to the MER framework, and validate their effectiveness on a moderately large-scale database.

The paper is organized as follows. We first present a novel ranking-based emotion annotation method in Section 2, and then describe automatic methods for ranking emotion in Section 3. The evaluation of the proposed annotation and prediction methods are reported in Section 4. Section 5 concludes the paper.

2. RANKING-BASED EMOTION ANNOTATION

For simplicity in execution and analysis, most existing psychological studies [6]–[8] adopted rating measure to help subjects track the emotion variation of a music piece as it unfolds in time. In our previous work [4] of building an MER system that views emotion from the continuous perspective and predicts the representative emotion of a song, we also adopted the rating measure and asked subjects to rate arousal and valence (separately) from –1.0 to 1.0 in 11 ordinal levels. However, in the course of the subjective evaluation, we found that the subjects had a prevalent cognitive difficulty in numerically rating the emotion to represent a song. This difficulty causes a serious user fatigue problem that largely reduces the reliability of the annotations, which in turn deteriorates the prediction accuracy of the MER system.

To address this issue, we propose to use a ranking measure for emotion and have the subjects make pairwise emotion comparisons. Since it is a lengthy process to annotate the straight ordering (which requires $n(n–1)/2$ comparisons for n music pieces), we propose a *music emotion tournament* scheme to reduce user fatigue. As shown in Fig. 2, n randomly chosen pieces are grouped in n–1 tournaments, which form a hierarchy of $\log_2 n$ levels. The results of the pairwise comparisons can then be incorporated to an $N \times N$ binary preference matrix P, with each entry (u, v) representing whether piece u is ranked higher than v, as exemplified in Fig. 2. N denotes the total number of music pieces in the database; usually $N >> n$.

We can then use the greedy algorithm proposed in [10] to efficiently approximate a global ordering π from the preference matrix P. The intuition is simple: the more items

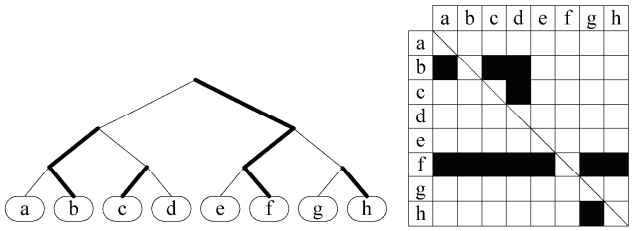

Fig. 2. Left: the proposed ranking-based emotion annotation method, which groups eight randomly chosen music pieces in seven tournaments. We use bold line to indicate the winner of each tournament. Right: the resulting preference matrix (partial), with the entry (u, v) painted black to indicate that piece u is ranked higher than v. The global ordering f>b>c=h>a=d=e=g can then be estimated by the algorithm described in Fig. 3.

Input: a list of data D and the associated preference matrix P
Output: an approximated optimal global ordering π
let $N=|D|$, $V=D$
for each $v \in V$, **do** $\rho(v)=\sum_{u \in V} P(v,u) - \sum_{u \in V} P(u,v)$
while V is non-empty **do**
  let $T=\arg\max_{u \in V} \rho(v)$
  for each $t \in T$ ***do*** $\pi(t)=N-|V|+1$
  $V=V-T$
  for each $v \in V$ and $t \in T$ **do** $\rho(v)=\rho(v)+P(t,v)-P(v,t)$
end while

Fig. 3. The greedy ordering algorithm, which is originally proposed in [10] and modified in this work to handle ties.

an item u dominates (ranked higher in pairwise comparisons), or the lesser items that u is dominated by, the greater ordering u would have. We have modified the algorithm to handle ties, which are present in our data because of large N. Due to space limitation, we simply list the pseudo codes in Fig. 3, and refer interested readers to [10] for more details.

3. RANKING-BASED EMOTION PREDICITON

To predict the AV values, existing approaches [3], [4] employ regression techniques [13], which aim at predicting a real value accurately. However, since MER is formulated as a ranking problem, we can also employ the *learning-to-rank* algorithms [11], [12] to directly optimize a ranking-based objective function for better accuracy. A schematic diagram of the training phase of this ranking-based MER system is shown in Fig. 4. The model training and 2DES mapping parts of the system are detailed below.

3.1. Learning-to-rank

The state-of-the-art methods fall into two categories: the pairwise [11] and the listwise approach [12]. The pairwise approach takes object pairs as learning instances, formulates the learning task as the classification of object pairs into two

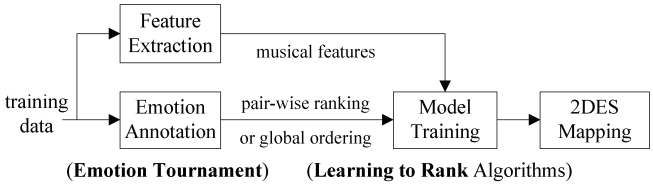

Fig. 4. The training phase of the ranking-based MER system.

categories (correctly and incorrectly ranked), and trains classification models for ranking. For example, in the seminal work of Herbrich et al [11], support vector machines are adapted to classify object pairs in consideration of large margin rank boundaries. Though the method, RankSVM, and its derivatives have been shown effective, they ignore the fact that ranking is a prediction task applied to a list of objects. Moreover, taking every possible pair is of complexity $O(N^2)$ and thus can be exceedingly time consuming.

The listwise approach conquers these shortcomings by using lists directly as learning instances and minimizing the *listwise loss* between the ground truth ranking list and the predicted one. For example, ListNet [12] employs linear neural network model and gradient descent techniques to minimize a probabilistic-based listwise loss function, and thus applies optimization directly on lists. Thanks to the linear kernel, the time complexity of ListNet is $O(N)$.

3.2. 2DES mapping

The outputs of learning-to-rank algorithms are the predicted orderings of music pieces, which are then mapped to the 2DES to generate a continuous representation. The music pieces which are ranked topmost (lowermost) are assigned with the maximal (minimal) arousal or valence values, and the remaining ones can then be mapped linearly or under some distribution. The AV values obtained by this normalization may be not as accurate (in terms of MSE) as the one predicted by regression models, yet as we have argued, perceptually human may place equal importance to the relative AV values to one another and the absolute AV values of each song. In addition, 2DES mapping is free from the semantic coverage problem of the regression-based methods.

4. EXPERIMENT

4.1. Experimental setup

The music database is composed of 1240 Chinese pop songs, whose emotions are annotated through the subjective test described in Section 4.2. Each song is on the average compared to 5.9 songs. Note the genre of our database is pop music rather than the western classical music as adopted in [1]–[3] since MER is to facilitate music retrieval and management and since it is the popular music that dominates the everyday music listening. The music pieces are converted to a uniform format (22,050 Hz, 16 bits, and mono channel PCM WAV) and normalized to the same volume level for fair comparison. Due to copyright issues, we use the 30-second segment starting from the 30th second of each song, a common practice in the field of high-level music classification [14].

We use Marsyas [15] and the MPEG-7 audio encoder [16] to extract 459 musical features, including Mel-frequency cepstral coefficients, spectral properties (centroid, moments, roughness, and crest factors), beat, harmonic ratio, and fundamental frequency type. It has been found that these features are correlated to music emotion [1], [4], [6].

Table I. Comparison of the rating- and ranking- based music emotion annotation through a large-scale subjective evaluation. The scores are in a five-point scale with '3' means neutral.

method	easiness	within-subject reliability	between-subject reliability
rating-based	2.82	2.92	2.81
ranking-based	4.07	3.78	3.36

4.2. Evaluation of the ranking-based annotation

To justify the feasibility of the proposed ranking-based annotation method, we design a web-based subjective test and invite subjects to annotate eight randomly selected music pieces using both rating- and ranking-based methods. For the rating measure, a scroll bar with end points denoting 0 and 100 is used. Since the prediction of arousal has been found relatively easy in previous MER work [1]–[4], we only ask subjects to annotate valence to reduce cognitive load. A questionnaire is presented at the end of the annotation process with the following three evaluation inquiries. All answers are on a five-point scale ranging from 1 to 5 (strongly disagree to strongly agree).

- **Easiness**. The annotation is easy to perform.
- **Within-subject reliability**. My annotation to the same songs would be nearly identical even after a month.
- **Between-subject reliability**. The annotation of others to the same songs would be nearly identical to mine.

A total of 602 subjects answer the questionnaire; the average results are tabulated in Table I. As the table shows, the ranking-based method is much easier to use than its counterpart. This validates our claim that it is easier for human to express a continuum of emotions in a comparative way. The subjects also express a high confidence level for the ranking-based annotation, either for within- or between-reliability. This implies that the ranking-based method is not only intuitive but also helpful to reduce the inconsistency of emotion annotation. Moreover, since all the results for the rating-based method are below the borderline, the necessity and importance of the proposed method is evident.

Table II. Comparison of different learning models in terms of Kendall's τ (for valence prediction) and execution time.

method	learning type	kernel	τ	time / iter
random	–	–	0.092	–
SVR	regression	RBF	0.401	2.8 sec
rankSVM	learning-to-rank	linear	0.335	~ 3 hr
ListNet	learning-to-rank	linear	0.428	1.0 sec

4.3. Evaluation of the ranking-based emotion prediction

We evaluate the prediction accuracy of emotion rankings in terms of Kendall's τ [17], the most frequently used statistical measure for comparing the ordinal correlation of two random variables. It is defined by the number C of concordant pairs (ranked correctly) and the number D of discordant pairs (inversions) as follows:

$$\tau = \frac{C - D}{C + D} = \frac{2C}{N(N-1)/2} - 1 . \quad (1)$$

τ has value 1 for perfect agreement, and –1 for total inverse agreement. We compare the performance of support vector regression (SVR) [13] and the two famous learning-to-rank algorithms, RankSVM [11] and ListNet [12]. All the three methods use the approximated global ordering π as input. The implementations of SVR and RankSVM are based on the free libraries LIBSVM [18] and SVM^{light} [19] with default parameters, respectively. ListNet is implemented in MATLAB. The programs are run on a regular Intel Pentium server. We randomly select one-fifth of data as test set and use the remaining ones for training. The average results are obtained by repeating the evaluation process 100 times.

Table II shows the experimental result. All the learning-based methods outperform the random permutation baseline by a great margin. The pair-wise approach RankSVM is less effective and much more time-consuming than its listwise counterpart, similar to the result reported in [12]. Among all, ListNet achieves the highest Kendall's τ, 0.428, though the performance difference between ListNet and SVR seems not significant. This is reasonable since by predicting real values a regression model is also solving a ranking problem.

5. CONCLUSION

In this paper, we have presented a novel ranking-based MER framework. The major contributions of this work are two-fold. First, we propose a music emotion tournament scheme and ask subject to annotate in a comparative way, which greatly reduces the cognitive load of annotation and enhances the reliability of training data. Second, we propose to formulate MER as a ranking problem and rank music by arousal or valence values. This corresponds to the intuition that human is also sensitive to the ordering of songs relative to one another besides the absolute emotion value of each song. With moderately large-scale subjective and objective evaluations, the feasibility of the ranking-based emotion annotation and prediction methods is verified; we obtain a high degree of easiness and reliability for emotion annotation, and a Kendall's τ up to 0.43 for rank prediction.

6. REFERENCES

[1] L. Lu et al, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.

[2] T. Li and M. Ogihara, "Content-based music similarity search and emotion detection," *Proc. ICASSP*, pp. 17–21, 2006.

[3] M. D. Korhonen et al, "Modeling emotional content of music using system identification," *IEEE Trans. Syst., Man., Cybern.*, vol. 36, no. 3, pp. 588–599, 2006.

[4] Y.-H. Yang et al, "A regression approach to music emotion recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.

[5] Y.-H. Yang et al, "Mr. Emo: Music retrieval in the emotion plane," *Proc. ACM MM*, 2008, accepted.

[6] E. Schubert, "Measurement and time series analysis of emotion in music," Ph.D. dissertation, School of Music Education, Univ. New South Wales, Sydney, Australia, 1999.

[7] J. A. Russell, "A circumplex model of affect", *Journal of Personality & Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[8] F. Nagel et al, "EMuJoy: Software for continuous measurement of perceived emotions in music," *Behavior Research Methods*, vol. 39, no. 2, pp. 283–290, 2007.

[9] S. Ovadia, "Ratings and rankings: Reconsidering the structure of values and their measurement," *Int. Journal of Social Research Methodology*, vol. 7, no. 5, pp. 403–414, 2004.

[10] W. W. Cohen et al, "Learning to order things," *Journal of Artificial Intelligence Research*, vol. 10, pp. 243–270, 1999.

[11] R. Herbrich et al, "Support vector learning for ordinal regression," *Proc. ICANN*, pp. 97–102, 1999.

[12] Z. Cao et al, "Learning to rank: from pairwise approach to listwise approach," *Proc. ICML*, pp. 129–136, 2007.

[13] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, pp. 199–222, 1998.

[14] N. Scaringella et al, "Automatic genre classification of music content: a survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, 2006.

[15] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, no.5, pp. 293–302, 2002. http://marsyas.sness.net/.

[16] MPEG-7 Audio Encoder. http://mpeg7audioenc.sourceforge.net/.

[17] M. Kendall, *Rank Correlation Methods*, Hafner, 1955.

[18] C.-C. Chang et al, "LIBSVM: a library for support vector machines," 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[19] T. Joachims, "Making large-scale SVM learning practical," *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.