

Personalized Music Emotion Recognition

Yi-Hsuan Yang, Yu-Ching Lin, and Homer Chen
National Taiwan University, Taipei, Taiwan R.O.C.

{affige, vagante}@gmail.com, homer@cc.ee.ntu.edu.tw

ABSTRACT

In recent years, there has been a dramatic proliferation of research on information retrieval based on highly subjective concepts such as emotion, preference and aesthetic. Such retrieval methods are fascinating but challenging since it is difficult to build a general retrieval model that performs equally well to everyone. In this paper, we propose two novel methods, bag-of-users model and residual modeling, to accommodate the individual differences for emotion-based music retrieval. The proposed methods are intuitive and generally applicable to other information retrieval tasks that involve subjective perception. Evaluation result shows the effectiveness of the proposed methods.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Retrieval Models; H.5.5 Sound and Music Computing: Systems, Modeling

General Terms: Algorithms, Performance, Human Factors

Keywords: personalization, emotion, perceptual residual

1. INTRODUCTION

Due to the explosive growth of digital data, innovative multimedia retrieval methods based on highly subjective concepts such as emotion [1], preference [2] and aesthetic [3] emerge as an alternative of conventional keyword-based methods. For example, a music emotion recognition (MER) system estimates the affective content of music signals so that a user can retrieve songs of certain desired emotions. Such a retrieval method is fascinating because it is content-centric and functionally powerful.

However, because human perception is intrinsically subjective, developing a general retrieval model that performs equally well for everyone can be very challenging. For example, each circle in Figure 1 corresponds to the annotation of a subject for a song over the 2D emotion plane, which defines emotion in terms of valence (how positive and negative) and arousal (how high and low) [4].¹ Evidently, simply assigning one emotion value to each song in a deterministic manner does not perform well in practice because the emotion perception varies greatly from person to person.

In this paper, two novel methods, bag-of-users model and residual modeling, are proposed to accommodate the individual differences in subjective emotion perception. A personalized MER system is built and evaluated to showcase the effectiveness of the proposed methods.

2. PROBLEM STATEMENT

We differentiate the following two MER problems:

- General recognition: Given a song s_i , recognize the emotion value y_i generally perceived by every user.

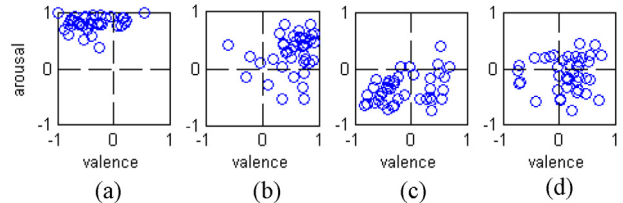


Figure 1. Emotion annotations in the 2D valence-arousal emotion plane [4] for four songs: (a) Smells like teen spirit by Nirvana, (b) A whole new world by Peabo Bryson and Regina Belle, (c) The rose by Janis Joplin, (d) Tell Laura I love her by Ritchie Valens. Each circle corresponds to the annotation of a subject.

- Personalized recognition: Given a song s_i and a user u_j , recognize the emotion value y_{ij} perceived by the user.

Despite that the subjective nature of emotion is well recognized, most previous works on MER only address the general recognition. In [1], emotion recognition is formulated as a regression problem. Given N inputs (\mathbf{x}_i, y_i) , $1 \leq i \leq N$, where \mathbf{x}_i is a feature vector of the i th song s_i , and $y_i \in [-1, 1]$ is the emotion value obtained by averaging the annotations of subjects, a general regression model $M(\cdot)$ is trained by minimizing the squared error between y_i and $M(\mathbf{x}_i)$, where $M(\mathbf{x}_i)$ is the recognition result for s_i . We use this method as the baseline in this paper.

3. PROPOSED METHODS

A schematic diagram of the personalized MER system is shown in Figure 2. To improve general recognition, we propose the bag-of-users model (BoU) to better utilize the annotations collected from subjective test. To make personalized recognition, we develop the residual modeling (RM) method based on the novel notion of *perception residual*. These methods are described in detail below.

3.1 Bag-of-Users Model

The ground truth data needed for training a retrieval model is typically obtained by averaging the opinions of subjects. This procedure, however, makes little use of the individual annotations assigned by each subject, which may provide abundant cues of the perception of a song. Figure 1(c) illustrates that simply averaging annotations loses the information that the human perception of the song is nearly bi-modal.

In the bag-of-users model, we train a regression model $M_j(\cdot)$ for each subject u_j using his/her annotations and obtain a bag of models $M_1(\cdot), M_2(\cdot), \dots, M_U(\cdot)$, where U denotes the number of subjects. We then aggregate the models using a super regression

¹ Valence and arousal are the two emotion dimensions found to be most fundamental by cognitive study [4]. The four quadrants of the valence-arousal emotion plane correspond to happy, angry, sad and relaxing.

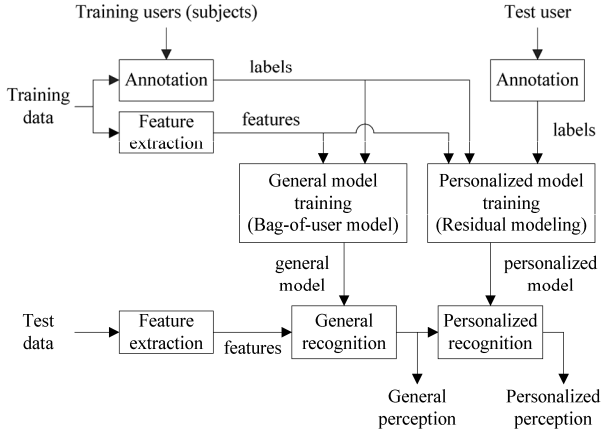


Figure 2. A schematic diagram of the personalized system.

model to make general recognition. Let $\hat{y}_i = [\hat{y}_{i1}, \hat{y}_{i2}, \dots, \hat{y}_{iU}]$ denotes a vector of the recognition results for s_i , where $\hat{y}_{ij} = M_j(\mathbf{x}_i)$. We train the super model $M^*(\cdot)$ by minimizing the error between y_i and $M^*(\hat{y}_i)$. The estimate $M^*(\hat{y}_i)$ can be regarded as the aggregation of the opinions of the U subjects.

3.2 Residual Modeling

Given the general perception y_i of a song s_i , we can compute the *perception residual* of a user u_j as the difference between the general perception and the personalized one, $r_{ij} = y_{ij} - y_i$. In this way, personalized recognition can be decomposed into the recognition of the general perception y_i and the perception residual r_{ij} . The former is more related to the content of the music sample, while the latter is more related to the user.

The recognition of perception residual is called residual modeling. We ask a test user u_z to annotate his/her perceived emotions of a number of songs Φ and use the annotations to train a personalized model $M_z(\cdot)$ that minimizes the error between r_{iz} and $M_z(\mathbf{x}_i)$, $i \in \Phi$. The personalized emotion is then computed by $y_{iz} = M_z(\mathbf{x}_i) + M^*(\hat{y}_i)$, if we use the bag-of-users model to recognize general perception.

4. EXPERIMENTAL RESULT

The music database consists of 60 popular songs from English albums. 99 subjects are recruited from the campus, making each song annotated by 40 subjects. Each song is represented by 80-dimension Mel-frequency cepstral coefficients—a widely-used feature representation for audio signal processing [5]. Support vector regression (SVR) is adopted to train the regression models. Our implementation of SVR is based on the library LIBSVM [6]. A standard measure of the goodness of fit for a regression model is the squared sample correlation coefficient R^2 between the ground truth and the estimated ones [7]. The value of R^2 ranges from 0 to 1; an R^2 of 1.0 means perfect fit.

For general recognition, we compare the performance of the bag-of-users model against the baseline model described in Section 2. We randomly select 10 songs as the test data and the remaining ones as training data. The overall procedure is repeated 1000 times to get the averaged result. While the R^2 of arousal is both around 0.70, the bag-of-users model improves the R^2 of valence from 0.149 to 0.158, a 6.4% relative improvement.

Table 1. Accuracy (R^2) of personalized valence recognition

Method	$ \Phi =5$	$ \Phi =10$	$ \Phi =20$	$ \Phi =30$
Baseline	0.1630	0.1635	0.1645	0.1639
BoU	0.1720*	0.1742*	0.1747*	0.1745*
RM	0.1632	0.1691	0.1768*	0.1839*
BoU+RM	0.1724*	0.1758*	0.1788*	0.1816*

*Significant improvement over the baseline at the $\alpha=0.01$ level

To evaluate personalized recognition, we use the 6 subjects who have annotated all the 60 songs as test users and the remaining ones as training users. The annotations of a test user u_z for $|\Phi|$ randomly selected songs are used to train a personalized model $M_z(\cdot)$, and the annotations of the same user for another 10 songs are used to evaluate $M_z(\cdot)$. The result of the personalized valence recognition with varying numbers of $|\Phi|$ is shown in Table 1, with each star denoting a significant improvement over the baseline model at the $\alpha=0.01$ significance level. It can be found that the proposed methods indeed improve personalized recognition. When $|\Phi|$ is small, the combination of BoU and RM performs the best; a 7.5% relative improvement is achieved with only $|\Phi|=10$ user inputs. Besides, the performance of RM increases as the number of $|\Phi|$ increases. When $|\Phi|$ is large, RM starts to outperform BoU+RM. A 12.2% relative improvement is obtained when $|\Phi|=30$. As for arousal, the baseline model and the personalized model produce similar results (about 0.65). This is not surprising since it has been pointed out that the perception of arousal is much less subjective than that of valence [1].

5. CONCLUSION

We have presented two methods to accommodate individual differences for subjective concept-based information retrieval systems. The bag-of-users model provides a better way to aggregate the individual perceptions of the subjects, while the residual modeling makes a personalized system focus on music content and user perception in different stages. The novel perspectives introduced in this paper can be applied to other applications that involve subjective human perception.

6. ACKNOWLEDGMENTS

This work was supported by the National Science Council of Taiwan under contract number NSC 97-2221-E-002-111-MY3.

7. REFERENCES

- [1] Y.-H. Yang et al, "A regression approach to music emotion recognition," IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 2, pp. 448–457, 2008.
- [2] A. S. Harpale and Y.-M. Yang, "Personalized active learning for collaborative filtering," ACM SIGIR, pp.259–266, 2008.
- [3] C. Dorai et al, "Computational media aesthetics: finding meaning beautiful," IEEE Multimedia, vol. 8, no. 4, 2001.
- [4] R. E. Thayer, The Biopsychology of Mood and Arousal, New York, Oxford University Press, 1989.
- [5] M. A. Casey et al, "Content-based music information retrieval: Current directions and future challenges," Proceedings of the IEEE, vol. 96, no. 4, pp. 668–696, 2008.
- [6] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001.
- [7] A. Sen and M. Srivastava, Regression Analysis: Theory, Methods, and Applications, New York, Springer, 1990.