Music Emotion Annotation by Machine Learning

Wai Ling Cheung and Guojun Lu

Gippsland School of Information Technology
Monash University, Churchill, Victoria 3842
Australia
{Wai.Ling.Cheung, guojun.lu}@infotech.monash.edu.au

***Abstract*—Music emotion annotation is a task of attaching emotional terms to musical works. As volume of online musical contents expands rapidly in recent years, demands for retrieval by emotion are emerging. Currently, literature on music retrieval using emotional terms is rare. Emotion annotated data are scarce in existing music databases because annotation is still a manual task. Automating music emotion annotation is an essential prerequisite to research in music retrieval by emotion, for without which even sophisticated retrieval methods may not be very useful in a data deficient environment. This paper describes a machine learning approach to annotate music using a large number of emotional terms. We also estimate the training data size requirements for a workable annotation system.**

Our empirical result shows that 1) the task of music emotion annotation could be modelled using machine learning techniques to support a large number of emotional terms, 2) the combination of sampling method and data-driven detection threshold is highly effective in optimizing the use of existing annotated data in training machine learning models, 3) synonymous relationships enhance the annotation performance and 4) the training data size requirement is within reach for a workable annotation system. Essentially, automatic music emotion annotation enables music retrieval by emotion to be performed as a text retrieval task.

I. Introduction

Most people experience emotions in music. It is a strong inner feelings triggered by musical expressions that could be spontaneously perceived by all, even young infants. Naturally, we describe emotional experiences using vocabularies in human language called emotional terms, such as *cheerful* and *gloomy*. The task of music emotion annotation is to attach emotional terms to musical works describing emotional experiences arise from music. This paper proposes to automate this manual annotation task using machine learning techniques.

Music emotion annotation is inherently different from existing works on mood classification in a number of ways. First of all, the purposes they serve are totally different. Mood classification aims to *separate* musical works and sort each of them into one emotional category, a one-to-one relationship, while emotion annotation aims to *describe* emotions in music using multiple emotional terms, a one-to-many relationship. For example, *Beethoven's moonlight sonata* is sorted into the *calming* category in mood classification; in emotion annotation the same musical work would be annotated as *romantic*, *loving* and *relaxing*. Secondly, emotion annotation involves a larger number of specific emotional terms as compared to a smaller number of general categories in mood classification. Current works [1] [2] [3] on mood classification mainly focus on classifying musical works into 2 to 8 categories. In this study our music collection contains 97 unique emotional terms in total, and our proposed annotation method is designed to accommodate 1,000 emotional terms and beyond.

Thirdly, the learning targets in mood classification and emotion annotation are different. In mood classification, emotional categories are mutually exclusive and they usually have very general meanings. In emotion annotation, emotional terms do not necessarily exclude one another, and they can have more specific meanings as compared to emotional categories.

In our study, we utilize musical excerpts downloaded from online music websites [4] [5], each comes with several emotional terms annotated by human subjects. Seventeen musical features are extracted from this music collection, which is adopted from existing mood classification research [6]. They are then fed to machine learners to discern their relationship with emotional terms. Once the machine learning models are trained, they are useful for the music emotion annotation task.

In the following subsections, we shall explain why automating music emotion annotation is necessary.

A. Access to Content Semantics

In the last decade, the size of music database has expanded to an unprecedented level. In the face of vast music database and its increasing volume, users from music industry inevitably demand for greater accessibility to musical contents. However, recent research mainly focuses on search and retrieval facilities for factual data such as *artist* and *title*. Query by emotion remains a largely unfulfilled part of music information retrieval systems, due to the present situation where most musical data is not annotated with emotional descriptions. In fact, query by emotion greatly enhances the usability of a music information retrieval system, as it facilitates search and retrieval by content semantics in addition to factual data.

Other researchers such as Huron [7] points out the limitations of factual reference information in musical queries and asserts that emotion is one of several most useful *retrieval indexes* of musical content. Factual data inherently exists with music and in many cases it is readily available. In contrary, music emotion is subject to listener's interpretations and data of this nature is very scarce in current music databases. Consequently, research on query by emotion is restrained by the lack of existing emotion annotated data. Downie describes the general lack of test data in [8]; he proposes the creation

978-1-4244-2295-1/08/$25.00 © 2008 IEEE

of *data-rich query records* as a music research repository which includes the descriptions of *mood*. To enrich research on query by emotion, more annotated works must first be made available.

B. Reduce Human Annotation Effort

At present, music emotion annotation is still a manual task by human annotators. Manual emotion annotation is tedious and costly, for each piece of music is listened to by several human subjects to ensure overall consistency. In a more stringent empirical setting, each subject is required to repeat the same listening task at different times to eliminate variations caused by external factors. Besides, the amount of time needed to annotate a multi-million song database is simply breathtaking. Under these circumstances, the current manual practice in music emotion annotation is unsustainable and research on automating this process is long overdue.

The rest of the paper is organized as follows: Section II reviews the current literature related to music emotion annotation. The proposed method is described in Section III. Experimental settings and results are given in Section IV. The results are then discussed in Section V, followed by our concluding notes in Section VI.

II. Related Works

One interesting aspect of this research is its highly multi-disciplinary nature involving musicology, psychology, information retrieval and machine learning. To our best knowledge, little or no research on automated music emotion annotation is available. In the following subsections, we shall first highlight the major works in mood classification, paying particular attention to their uses of emotion taxonomies from musicology and psychology, as well as their adopted machine learning techniques. Based on the recent development of these disciplines, the second subsection explores the technical challenges in annotating emotion terms to musical works.

A. Emotional Taxonomies, Musical Features and Machine Learning Techniques in Mood Classification

In 2003, emotions are classified into 13 categories using 499 samples in [9]; however, the categories are later regrouped to six due to unsatisfactory result. From Farnsworth's revision [10] of Henver's checklist [11] that suggests ten categories, the authors add three extra categories to form the initial 13 emotional categories. Thirty musical features on timbral texture, rhythmic content and pitch content are extracted from MARSYAS [12]. Musical works are classifies into multiple emotional categories using a set of binary classification problems by Support Vector Machine. The authors admit that the low performance is mainly attributed to the misclassification of borderline cases. When they scale back to six categories, the result is more acceptable.

Feng et al. [2] classifies emotions on 4 categories adopted from Juslin's theory [13] of basic emotions using 353 pieces of pop music. Tempo and articulation are used with Neural Networks to classify musical works into one of four emotional categories. The classification result is satisfactory, but only 23 samples are used in the testing set.

Yang et al. [14] classifies 500 songs into 2 emotional categories, "positive" or "negative' emotions. Emotion intensity, tempo, timbrel features and information from text lyrics are used to train Support Vector Machine Regression models to classify musical works. The experimental result is better than other studies but this 2-category problem is also simpler than those with 4-8 categories, hence their results cannot be compared directly.

In 2006, Yang et al. [3] classifies 4 emotions of 195 audio segments based on Thayer's [15] model of four quadrants as four emotional categories. Fifteen musical features on loudness, pitch and duration are adopted from Schubert's recommendation [16]. This implementation uses fuzzy KNN and Fuzzy Nearest-Mean to handle borderline cases lying between emotional categories. Later Yang et al. [17] transform the emotional representation from categories onto a 2-dimensional plane using regression models of Multiple Linear Regression, Support Vector Machine and AdaBoost. The experimental results are impressive, but less than 20 samples are used in the testing set.

In 2007, Chua [6] classifies 1011 songs into 8 emotional categories using the original eight categories from Henver's checklist [11]. Using Support Vector Machine classifier and its regression counterpart, seventeen musical features from tempo, articulation, rhythm, loudness, pitch and harmony are recommended by Chua. The results are satisfactory considering this is an 8-category problem which is more difficult than those attempted by other predecessors in mood classification.

B. Technical Challenges of Annotating Emotional Terms to Musical Works

One challenge raised by a number of research studies in mood classification [1] [2] [17] is the lack of uniformity in emotion taxonomies from musicology and psychology. As described in Section II-A, individual researchers define their own sets of learning targets derived from predefined emotional word list such as Henver's checklist [11] or emotions from Thayer's model [15]. Predefined word list is a guided form of description that restricts the choice of words listener used to describe music. By nature, guided descriptions are suitable for mood classification because they provide overall uniformity for classifying emotions. However, music emotion annotation requires emotional description that allow flexibility and subtle precision. Thus, rather than using a predefined word list from existing emotion taxonomy, we use emotional terms as *free* descriptions originated from listeners that are similar to emotional keywords in a sentence.

Annotating emotional terms as *free* description involves a higher number of emotional terms, which presents a challenge to the capacity of machine learning models due to an increase of problem complexity. To put this in perspective, Whissell's dictionary of affect [18] estimates that there are 4,000 emotion-loaded terms in modern English vocabulary. If we assume

one in four emotional terms is relevant to music, a realistic estimation of emotional terms for our proposed annotation method would be 1,000. Based on our preliminary experiment, drastically increasing the number of learning targets in existing mood classification model yields poor result. We therefore redefine this problem in Section III-B to accommodate a large number of emotional terms in the machine learning process, by dividing the emotion annotation task into a set of smaller sub-tasks.

Another challenge with annotation using emotional terms is imbalanced data. In most cases, fewer songs in a music database are annotated with an emotional term (positive samples) as compared to those without the annotation (negative samples). When positive samples are under-represented in the dataset, machine learning models are unable to learn emotional terms correctly which lead to poor annotation result. For this reason, sampling methods are introduced in Section III-C to balance positive/negative sample ratio.

We note that misclassification on borderline cases has been an area of concern in mood classification [9] [3] [6], and contemporary psychology research [19] also suggests emotional concepts tend to be fuzzy in nature. Thus, classification confidence is used in Section III-B to estimate the degree of emotion detected, in order to determine whether an emotion is present or absent. This estimation provides an opportunity to adjust the annotation sensitivity for very specific emotional terms, for their distribution is highly imbalanced. In Section III-D, we use a data-driven threshold to adjust annotation sensitivity in order to improve annotation performance on specific emotional terms.

Finally, human annotators tends to exclude synonyms in emotional description. Machine learning models trained with these manual annotation do not take synonymous relationships between emotional terms into account, although synonyms are also correct annotation in semantic sense, e.g. annotating *happy* to a *cheerful* song is semantically correct. Hence, we propose the use of thesaurus in Section III-E to annotate the meanings of emotional concept rather than simple term matching.

III. Method

In this section we shall discuss the music emotion annotation task at machine learning level. Section III-A provides a general overview of our proposed emotion annotation method using machine learning techniques. The subsequent subsections focus more specifically on problem definition, sampling methods, data-driven threshold and the use of thesaurus.

A. General Overview on Proposed Method

The machine learning process for our proposed method is illustrated in Figure 1. Firstly, excerpts of musical works are downloaded from online music websites [4] [5] as training and testing data. Seventeen musical features are then extracted from the excerpts to build machine learning models, they are listed as follows: 1) perceptual tempo, 2) perceptual tempo variation, 3) rhythm motion, 4) rhythm motion variation, 5) articulation, 6) articulation variation, 7) relative number of audible frequency sub-bands, 8) variation in relative number of audible frequency sub-bands, 9) roughness, 10) roughness variation, 11) relative perceived intensity of low frequencies, 12) variation in relative perceived intensity of low frequencies, 13) harmonicity, 14) harmonicity variation, 15) spectral flux, 16) spectral flux variation and 17) loudness. This set of musical features are directly adopted from Chua's study [6] as they are highly related to music emotion.

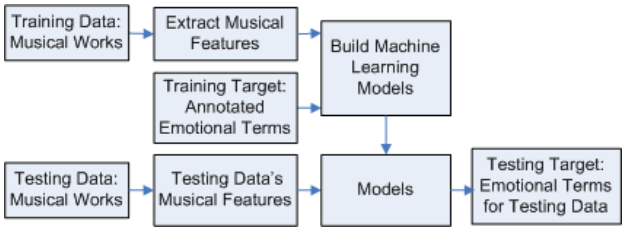

Fig. 1. Proposed Music Emotion Annotation using Machine Learning

Meanwhile, emotional terms annotated by human subjects for the excerpts are also downloaded as the training targets for model building. Once the machine learning models are built, musical features extracted from testing data are fed into models to annotate emotional terms. Finally, machine generated emotional terms are compared with manual annotated emotional terms (which is downloaded with testing data as testing targets) to determine annotation performance. To assess the generalization capability of our annotation method, *10-fold stratified cross validation* is used in all our experiments to evaluate model performance. We shall further explain our evaluation measures in Section IV.

The data set in our experiment contains 1103 excerpts from a wide range of instrumental and vocal music in various genres. This collection is the largest data set among existing works in Section II-A and it provides ample testing set (331 samples) after a 70/30 train/test split. Based on our preliminary tests on a number of learning algorithms using WEKA [20] data mining tool, we find that decision tree ensemble *Random Forests* [21] works well with the current data set. Its model building process resembles human decision making that is intuitive as compared to other learning algorithms, and the implementation is readily available for others to repeat the experiments. We shall use *Random Forests* in our experiments.

B. Problem Definition: Multiple 2-Class Problems

In Section II-B, we highlight the fact that annotation using emotional terms as *free* description involves a higher number of learning targets (1,000+) than mood classification, which drastically increase the problem complexity at machine learning level as compared to the usual n-class ($n \leq 8$) classification problem in mood classification. However, the learning targets of emotion annotation are also different from those in mood classification: emotion annotation uses multiple emotional terms per song, a one-to-many relationship, and emotional terms are not mutually exclusive to one another.

This one-to-many, non-exclusive relationship between emotional terms means that emotion annotation problem is *not* necessarily a single machine learning task. Instead of using one machine learning model to learn all emotional terms, the annotation task could be subdivided as a collection of separate sub-tasks using multiple models, with each model learns independently on one emotional term. Since each sub-task is independent from others, increasing the number of learning targets (emotional terms) does not affect problem complexity of individual sub-task. This annotation method therefore has the potential to accommodate a large number of emotional terms in the machine learning process.

In our proposed annotation method, the emotion annotation problem is recast into multiple 2-class problems, one for each emotional term with its own data set containing two class labels: 0 (absent) or 1 (present). A 2-class problem in machine learning implies a fixed class boundary within an emotional concept and the ground truth is two discrete class labels e.g. *sad* or *not sad*. However, emotional concepts tend to be fuzzy in nature [19], so it is more appropriate for the classifier to rank the samples according to the degree of membership of an emotional term. This degree of membership is typically called classification confidence in machine learning classification.

The estimation of classification confidence can be presented as a probability measure and it can be used as a way to adjust the decision threshold using cost-sensitive learning. The cost here is associated with the population of positive cases: the fewer the positive cases, the higher the cost. The emotion annotation problem is presented as multiple 2-class classification problems with each data set splits into training/testing sets: training set feed into an estimator to produce classification model which estimates classification confidence representing degree of presence of a particular emotion. We introduce a data-driven threshold to decide on the final class label. This shall be explored further in Section III-D - Data-Driven Threshold.

C. Sampling Method

Emotion annotation performance depends largely on data representation in the training data set. When imbalanced data set is presented with very few positive samples (samples with the emotional term we aim to annotate), machine learning model is unable to learn these emotional terms correctly which results in poor annotation performance. After the annotation problem is recast into multiple 2-class problems, the new data sets become dominated by the negative class (samples without emotional term - class "0") over the positive class (samples with emotional term - class "1"). For example, the *romantic* data set contains 71 positive samples and 1,032 negative samples, a positive-negative class ratio of 1:15. The data structure therefore needs to be re-balanced by sampling methods to better represent the positive class.

There are three commonly used sampling methods for imbalanced data sets as outlined in [22]: 1) Over sampling, where the minority class is repeated until its sample size is equal to the majority class; 2) Under sampling, where the majority class is reduced to a level that is closer or equivalent to the minority class (50% in our experiment); 3) Hybrid sampling, where the majority class is reduced to a pre-defined level (50% in our experiment), then the minority class is repeated until the number of samples in both classes is equal. We shall find out which sampling method is most effective in Section IV - Experimental Results.

D. Data-Driven Threshold

Sampling methods are more effective in less imbalanced data set since the number of positive samples is still sufficient to be used repeatedly to increase the positive class size. For highly imbalanced data set, the idea of data-driven threshold is introduced to improve annotation performance. In a 2-class problem, the normal threshold is usually fixed at 0.5 on positive class confidence to determine whether an emotion is *present* or *absent*. When the annotation problem is recast into multiple 2-class problems in Section III-B, we use the classification confidence to provide an opportunity to adjust annotation sensitivity by resetting the decision threshold for class boundary using the ratio of positive and negative cases in training samples.

We make use of a data-driven threshold k as a *flexible class boundary*: when more positive samples are available, the value of k is set higher to adjust this *flexible class boundary* so that the annotation precision is high; otherwise, the k value is lowered so that emotional terms with fewer positive samples could also be annotated successfully. By individualizing the threshold k for each emotional term, machine learning models could be adjusted by calculating k, where $k = \frac{|pos|}{|neg|}$ with 0.1 as the minimum value. Hence, emotional terms with less positive samples have lower threshold while general terms with more positive samples have higher threshold.

E. Use of Thesaurus

Human annotators tend to exclude similar terms in emotion description. For example, *cheerful* and *happy* are not usually annotated together; only one term is chosen because they are synonyms. But in query by emotion, songs annotated as *cheerful* are also relevant to a query on *happy* emotion. Machine learning models that are trained solely on data by human annotators do not take these synonymous relationships between emotional terms into account. For this reason, we propose to include synonyms as virtual members of the positive sample population in model training process to understand their effect on annotation performance. A synonym list between emotional terms is generated from online website [23], and synonyms of emotional terms are changed to positive in the training data sets.

IV. Experimental Results

A. Evaluation

Three performance indicators are used in our experiments to evaluate annotation performance: recall, precision and F-measure. Recall and precision are commonly used in information retrieval; Downie [8] describes them as

performance measures in music information retrieval and an example from mood classification could be found in [2]. Since recall and precision tend to work in reciprocal ways, a high reading in either recall or precision does not necessary mean good annotation performance; we need a combined indicator of recall and precision to compare the performance between different models. F-measure is the harmonic mean of recall and precision which provides a balanced evaluation of the former indicators. It is also used in mood classification [9]. The performance indicators are calculated as follows:

$$\textit{Recall} = \frac{No.\ of\ correctly\ annotated\ cases\ for\ a\ term}{Total\ number\ of\ positive\ cases\ with\ this\ term}$$

$$\textit{Precision} = \frac{Number\ of\ correctly\ annotated\ cases\ for\ a\ term}{Total\ number\ of\ cases\ annotated\ with\ this\ term}$$

$$\textit{F-measure} = \frac{(2 \times recall \times precision)}{(recall + precision)}$$

All three measures are obtained by *10-fold stratified cross validation*. Under normal 10-fold cross validation without stratification, the data set is randomly split into train/test sets for model generation and performance evaluation, and this process is repeated ten times before the results are averaged. However, random train/test split of the entire data set could result in the absence of positive samples in testing set; in this case performance of machine learning model would be overstated, e.g. performance readings are very high when the machine learning models are only good at omitting negative samples from annotation. In order to ensure results are not misleading, it is necessary to apply 10-fold stratified cross validation, in which training and testing sets are sampled independently from positive and negative samples in order to retain their positive/negative sample distribution after the train/test split.

In the case of semantic annotation, for a particular emotional term, correctly annotated cases are 1) if a model returns a positive case where the ground truth is marked as positive or 2) if a model returns a positive case where at least one synonym of this term is being marked as positive in the ground truth. This treatment is applied to all results in this paper.

B. Results

We present two models for comparison in Table I:

1) ETA (**E**motional **T**erm **A**nnotation): a base model that detects the presence of emotion by multiple 2-class machine learning models.
2) THETA (using **TH**esaurus in **E**motional **T**erm **A**nnotation): an ETA model that annotates with the use of hybrid sampling, data-driven threshold and synonymous relationships.

Since annotation performance is closely related to the number of positive training samples available to machine learning models, the experimental results in Table I are arranged in four tiers (T1-T4) to represent the average performance of all terms within the tier. Where $|pos|$ denotes the number of positive samples, T1 is $120 \leq |pos|$; T2 is $70 \leq |pos| < 120$; T3 is $20 \leq |pos| < 70$; T4 is $|pos| < 20$.

TABLE I
RECALL, PRECISION AND F-MEASURE FOR ETA AND THETA. BOLDFACED ARE BETTER F-MEASURE.

	Recall		Precision		F-measure	
	ETA	**THETA**	**ETA**	**THETA**	**ETA**	**THETA**
T1	0.14	0.59	0.79	0.67	0.24	**0.62**
T2	0.05	0.65	0.59	0.42	0.09	**0.49**
T3	0.02	0.51	0.27	0.33	0.03	**0.35**
T4	0.00	0.31	0.05	0.16	0.00	**0.18**

Our experimental result in Table I shows a significant improvement in F-measure from ETA to THETA across all tiers (T1:2.6 times; T2:5.4 times; T3:11.7 times; T4:18 times). It shows that hybrid sampling, data-driven threshold and the use of synonyms work together effectively to enhance ETA's annotation performance, which results in 0.59 (recall), 0.67 (precision) and 0.62 (f-measure) for T1 in THETA. There is no directly comparable results available in the music information retrieval field since mood classification is inherently a different problem, it would be inappropriate to compare its results with annotation. But these values in annotation performance are considered high as compared to a similar text retrieval system using synonyms, which is around the 60% mark [24].

V. DISCUSSION

A. To Support a Large Number of Emotional Terms

From Section II-B, the emotional terms used in our annotation system are *free* description that involves a large number of emotional terms. To achieve this, the annotation task is cast into multiple independent sub-tasks in Section III-B so that the annotation system would be capable of accommodating 1000+ emotional terms.

In our experiment, the speed of model building and evaluation is approximately 7.2 seconds per term using a 2.8GHz computer with 1G memory. In reality model evaluation is no longer required in annotation, thus build a model should take less than 7.2 seconds. Whissell's Dictionary of Affect [18] suggests there are 4000 emotional words in English vocabularies. Let us take 7.2 seconds per model as an indication; then building 1,000 models for 1,000 emotional terms should only take two hours. Even if we include all 4,000 terms from Dictionary of Affect in annotation, building 4,000 models should only take about eight hours which is still quite achievable.

B. Sampling Methods and Data-Driven Threshold

After recasting the annotation problem into multiple 2-class problems, the distribution of class labels becomes highly imbalanced in ETA. As a result, 67 terms (69%) score zero for F-measure in the experiment which means these emotional concepts could not be annotated by ETA. The low recall in T2-T4 indicates that ETA is only able to annotate at most 5% of songs from the data set for emotional terms with fewer than 120 positive training samples. We therefore introduce a sampling method and a data-driven threshold to increase the recall rates. In the experiment of ETA with different

sampling methods, we find that applying sampling methods generally improve recall rates. In particulr, hybrid sampling scores better F-measure than other sampling methods across all tiers, which is consistent with the findings in [22]. When data-driven threshold is added to ETA hybrid sampling, the recall rate increases further in all tiers. We note that precision in T1 and T2 compromised slightly, but the overall performance indicator f-measure still records significant increase due to the marked improvement in recall. The number of unannotable terms is also reduced from 67 to 6.

C. The Use of Thesaurus in Annotation

Generally, the use of synonym in training data set has a positive impact on annotation performance, especially for emotional terms that have some synonymous relationships with others and very few positive training samples. For emotional terms with less than 20 positive samples (i.e. T4), there is usually not enough samples for the machine learners to learn from. Using synonyms as virtual members to the positive sample population provides more information to the machine learning models to correctly identify a term. There are a few terms in THETA that could not be annotated. Since these terms have only one or no synonymous relationship with other terms and fewer than 10 positive training samples, they cannot be learned by machine learning models.

D. To Estimate Training Data Size Requirements for a Workable Annotation System

It is obvious from our experimental results that the more positive training samples, the better the annotation performance. For an annotation performance at approximately 50%, 70 (9% from 772 training samples) or more positive training samples are required. For a high annotation performance at 60%, approximately 120 (15.5% from 772 training samples) positive training samples are needed in the data set. The percentage provides an indication to the positive training samples size in our experiment; however, once THETA is given sufficient number of positive training samples to learn the emotional concepts (i.e. to draw a decision boundary between present or absent of an emotion), annotation of these concepts to musical works could be applied to different data sets. The experimental results convey a very positive message to music emotion annotation by machine learning, as the required positive training sample size is within reach to enable the automation of this annotation task.

VI. Conclusion

This research advances the music emotion annotation problem from a stage of manual operation to a stage of automation. The novelty of this research lies in the multidisciplinary nature of music emotion annotation. It automates a traditionally manual annotation task using a number of techniques from various disciplines, which is highly original. We contribute to show that automatic music emotion annotation is possible and workable using hybrid sampling, data-driven detection threshold and synonymous relationships between emotional terms, in conjunction with state-of-the-art machine learners. Our empirical result shows that training data size requirement is within reach for a workable annotation system. As music emotion description becomes readily available through automatic annotation, the development of a music research repository will be more attainable. Music retrieval by emotion will be a simple text retrieval task performed on emotion annotated music.

VII. Acknowledgements

The authors would like to thank B. Y. Chua for kindly making her research data available.

References

[1] D. Liu, L. Lu, and H. Zhang, "Automatic mood detection from acoustic music data." in *ISMIR*, Baltimore, Maryland, USA, 2003.
[2] Y. Feng, Y. Zhuang, and Y. Pan, "Music information retrieval by detecting mood via computational media aesthetics," in *IEEE/WIC International Conference on Web Intelligence*, Halifax, Canada, 2003.
[3] Y. Yang, C. Liu, and H. Chen, "Music emotion classification: a fuzzy approach," in *ACM Multimedia*, Santa Barbara, California, USA, 2006.
[4] "All music guide." [Online]. Available: http://wm03.allmusic.com/
[5] "Song peddler." [Online]. Available: http://www.songpeddler.com/
[6] B. Y. Chua, "Automatic extraction of perceptual features and categorization of music emotional expression from polyphonic music audio signals," Ph.D. dissertation, Monash University, 2007.
[7] D. Huron, "Perceptual and cognitive applications in music information retrieval." in *International Symposium on Music Information Retrieval*, 2000.
[8] J. Downie, "Toward the scientific evaluation of music information retrieval systems," in *ISMIR*, Baltimore, Maryland, USA, 2003.
[9] T. Li and M. Ogihara, "Detecting emotion in music," in *ISMIR*, Baltimore, Maryland, USA, 2003.
[10] P. R. Farnsworth, "A study of the henver's adjective circle," *Journal of Aesthetics and Art Criticism*, vol. 13, pp. 97–103, 1954.
[11] K. Henver, "Experimental studies of the elements of expression in music," *American Journal of Psychology*, vol. 48, pp. 246–268, 1936.
[12] G. Tzanetakis and P. Cook, "Marsyas: A framework for audio analysis," *Organized Sound*, vol. 4, no. 3, pp. 169–175, 2000.
[13] P. N. Juslin and J. A. Sloboda, *Music and emotion: theory and research*. New York: Oxford University Press, 2001.
[14] D. Yang and W. Lee, "Disambiguating music emotion using software agents," in *ISMIR*, 2004, pp. 52–58.
[15] J. F. Thayer, "Multiple indicators of affective response to music." Ph.D. dissertation, New York University, New York, 1986.
[16] E. Schubert, "Measurement and time series analysis of emotion in music," Ph.D. dissertation, University of New South Wales, Sydney, 1999.
[17] Y. Yang, Y. Lin, Y. Su, and H. Chen, "Music emotion classification: a regression approach," in *IEEE Interntational Conference on Multimedia and Expo*, 2007.
[18] C. M. Whissell, *Emotion: theory research and experience*. New York: Academic Press, 1989, vol. 4, ch. The dictionary of affect in language, pp. 113–131.
[19] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
[20] I. Witten and E. Frank, *Data mining: practical machine learning tools and techniques with JAVA implementations.* USA: Academic Press, 2000.
[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[22] Y. Liu, A. An, and X. Huang, "Boosting prediction accuracy on imbalanced datasets with svm ensembles," in *PAKDD 2006*, 2006.
[23] "Thesaurus.com." [Online]. Available: http://thesaurus.reference.com/
[24] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran, "Indexing with wordnet synsets can improve text retrieval," in *COLING/ACL'98 Workshop on Usage of WordNet for NLP*, Montreal, 1998.