Music Information Retrieval by Detecting Mood via Computational Media Aesthetics

Yazhong Feng Yueting Zhuang Yunhe Pan
College of Computer Science, Zhejiang University, China
fengyz_zju@263.net yzhuang@cs.zju.edu.cn yhpan@sun.zju.edu.cn

Abstract

It is well known that music can convey emotion and modulate mood, to retrieval music by mood is sometimes the exclusive manner people select music to enjoy. This paper concentrates on music retrieval by detecting mood. Mood detection is implemented on the viewpoint of Computational Media Aesthetics, that is, by analyzing two music dimensions, tempo and articulation, in the procedure of making music, we derive four categories of mood, happiness, anger, sadness and fear. Concretely, with regard to music in the format of raw audio, after tempo is detected using a multiple-agent approach, a feature called relative tempo is calculated, and after the mean and standard deviation of the feature called average silence ratio in the presented computational articulation model are calculated, a simple BP neural network classifier is trained to detect mood. Users retrieval music by browsing the 3D visualization of feature space associated with specific mood. This paper reports the experimental result on a test corpus of 353 pieces of popular music with various genres.

1. Introduction

With the booming of digital music on the Internet, music consumption grows to an unprecedented volume for its least cost and easily availability, the activity of music retrieval is now the heavy traffic on the Internet, music retrieval system and music recommender system are becoming eagerly needed software tools on Web.

In the community of music information retrieval, researchers developed methods to retrieval music with a particular melody, the queries were formulated by humming or singing [1][2][3][4][5]; they also developed methods to retrieval music by similarity [6][7][8]. The overviews about music information retrieval can be found in [9][10][11].

Music recommender system serves to provide personalized music contents such as a playlist that caters to the user's music taste or purchasing list that meets the user's need or preference, its popular technique is collaborative filtering, which bases its recommendation on the rating of near neighbors [12], but hybrid systems with other technologies such as content-based, utility-based, knowledge-based and demographic [13] are also paid more attentions, we think deriving features from music itself is extremely important for recommend new music in addition to modeling user patterns.

The most appealing function of music is that it can convey emotion and modulate listener's mood. The mystery of this function is well discussed in the field of psychology, cognition and aesthetics; nevertheless, automatically detecting mood is still in its infancy. We aim to retrieval music by mood, which is sometimes the exclusive manner people select music to enjoy, for example, when someone is sad for some reason, she or he wants to listen to a piece of music that can cheer her or him up, at this moment she or he will search music by mood no matter what the melody sounds or which the piece of music is similar to.

Unfortunately, to our knowledge, few system claims to be able to automatically retrieval music by mood, the possible reason is that at least two obstacles lie on this approach that make it very difficult, one is that there is no computational mood model yet, and the other is that mood is an item related with cultural background and involved scenario. In the essence, the difficulty stems from the gap between the rich meaning that users want when they query and the shallowness of content descriptor we can actually compute. Recently, a new approach, Computational Media Aesthetics (CMA) is presented aiming to fill this gap, it suggests to analysis media content on the basis of understanding compositional and aesthetics media principles. The core trait of it is that data is interpreted with its maker's eye [14]. Some works [15][16][17] extracting video semantics on the viewpoint of Computational Media Aesthetics are reported, with most of them using film grammar as the underpinning, but few work contributes to music retrieval based on the viewpoint of CMA. Composers choreograph the expectation to arise emotion, and performers convert the musical intention into music language to arise emotion, which inspires us to analysis music mood on the viewpoint of how music is made, i.e. the viewpoint of CMA avoiding coping with the ambiguity of emotion audience arise when confronted

Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03)
0-7695-1932-6/03 $17.00 © 2003 IEEE

with music. The proposed technique will be also useful for music recommender system.

The rest of this paper is arranged as followings: section 2 summarizes the related work on music mood detection; section 3 gives the overview of our scheme, section 3 describes how music tempo is detected; section 4 presents a computational articulation model; section 5 is the implement of mood detection; section 6 briefly introduce the user interface; section 7 gives the experimental results; section 8 concludes this paper and discusses the future research directions.

2. Related work

In Webster's dictionary, music mood is "a state of mind in which an emotion or set of emotions gains ascendancy", we do not differentiate between mood and emotion in this paper. Because there is few computational music mood model presented in the literature of music retrieval, we convey the literature of emotion detection in speech, which we think relates our work to some extent. [18] reports a pilot work on detecting emotion in speech, global statistics of pitch information and their variations are used as two sets of prosodic features to recognize emotion, four emotion labels, the same as what we label music mood in this paper, are classified by several statistical pattern recognition techniques, for its testing corpus of over 1000 utterances, automatically recognized emotion is compared with human labeled one, the error rate is about 10% larger. [19] uses realistic data from Wizard of Oz (WoZ) scenario to distinguish between two classes of emotion, anger and neural, different feature sets, prosodic and lexical, and multi layer perceptron classifier are employed, the average recall and precision rates are all about 60%.

Because retrospection of mood links with genre preference, music genre classification is also related with music retrieval by detecting mood, in [20], three feature sets for representing timbral texture, rhythmic content and pitch content, specifically spectral centroid, spectral rolloff, spectral flux, time domain zero crossings and mel-frequency cepstral coefficients (MFCC) are proposed to detect music genre.

Mood detection is the predominant task of our approach for music retrieval; to fulfill this task, we will also derive features from music audio signal as mentioned above in related work and then map feature space to mood space. As for the taxonomy of music mood, different clusters were suggested [21], in this preliminary work, to simplify the mood detecting procedure, we classify mood into four categories, which will be described in details in next section.

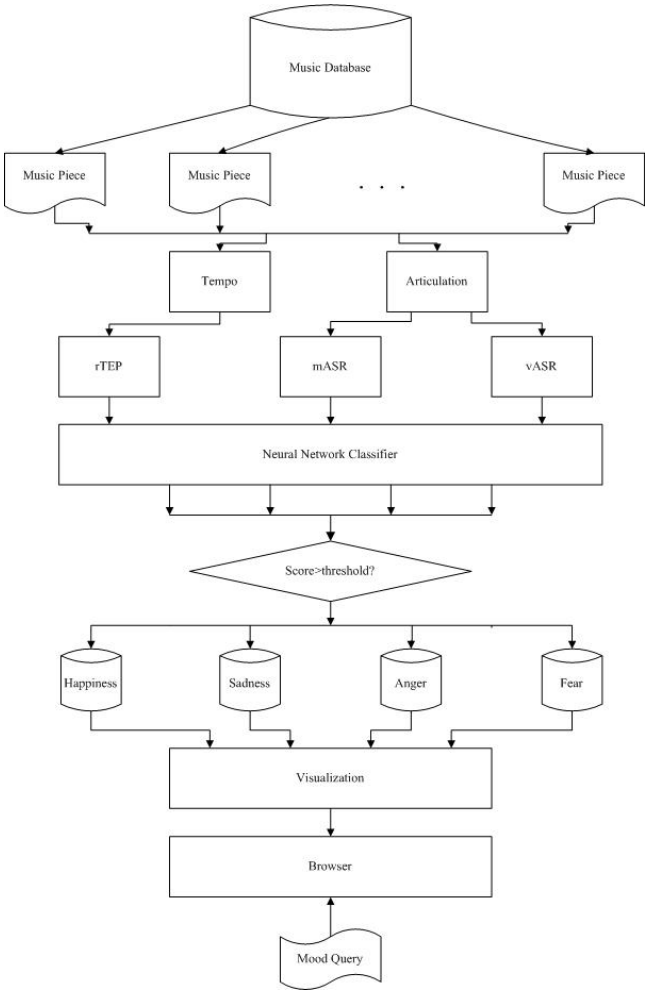

Figure 1. Overview of music retrieval approach by detecting mood.

3. Overview

Ideal music retrieval system should provide accesses to music in various formats via any indexing method. But to test our approach, we concentrate on the scheme of raw audio music retrieval by its mood dimension, which could act as a component of more comprehensive system if effective. Figure 1 gives the overview of our scheme. Music database is indexed on four labels of music mood, concretely, "happiness", "sadness", "anger" and "fear"; three features, relative tempo, the mean and standard deviation of average silence ratio, are used to classify mood, the classifier is a BP neural network. When user's query about music mood is accepted, the system displays the corresponding region of music database by visualizing the feature space of the music pieces, users then browse the space to select music piece.

Table 1. The relationship of music mood with tempo and articulation

MOOD	TEMPO		ARTICULATION	
	fast	slow	staccato	legato
happiness	yes	no	yes	no
sadness	no	yes	no	yes
anger	yes	no	no	yes
fear	no	yes	yes	no

Though we already have Thayer's model of mood [22], the theory of mood management by Zillmann [23], we find Juslin's theory is more appropriate for automatically detecting music mood. Inspired by CMA methodology, we will experiment on mood detection by analyzing music-making procedure, Juslin found that two music dimensions could explain the transfer of emotional content from performer to audience: tempo and articulation, tempos were either fast or slow while articulations were either staccato (very brief notes, separated by brief silences) or legato (playing so that one note glides into the next without interruption) [24], Table 1 explains how four kinds of typical mood are related with tempo and articulation.

The hypothesis we make in our music mood detection scheme is that tempo and articulation are invariable in the concerned music segment, though it is not uncommon in music, especially classical music, that fluctuations of tempo and articulation are present (future work will cover this situation). The variation of music tempo and articulation reveals the movement of music mood.

4. Tempo detection

Tempo is the rate at which notes are played, expressed in score time units per real time unit, i.e. quarter notes per minute or beats per minute , we define a real-valued interval $[s, f]$ representing the tempo boundary of most music pieces, s is the lowest tempo and f is the fastest tempo, s and f are statistically derived in section 8.2, a feature called relative tempo is defined on this interval:

$$rTEP = \frac{1}{(s+f)TEP} \quad (1)$$

Where TEP is the detected tempo by the tempo detection approach.

Music tempo detection is well studied in the communities of computer music and signal processing; some good algorithms are published [25][26][27][28]. The approach in [25] does not use any priori knowledge such as style, time signature or approximate tempo about music, nor does it model the cognition mechanisms involving human rhythm perception. It is robust in various music styles and computation coat is moderate. Deducted tempo is expressed in seconds, concretely *inter-beat interval* in this approach.

5. Computational articulation model

There are two ways of articulating successive tone during music performance, legato and staccato. When one tone is terminated near the onset point of the following tone, the tones are perceived as connected or legato, but when the first tone is terminated before the onset of the following tone, there are noticeable silences between the tones, they are perceived as staccato. We propose to model articulation by detecting the silence in the music signal, the silence is measured according to the energy of audio signal in time domain.

5.1. Feature extraction

A time-domain feature called Average Silence Ratio (ASR) is employed to model articulation, whose definition is:

$$ASR = \frac{1}{2N}\sum_{n=0}^{N-1}\left(1-\mathrm{sgn}\left(STE(n)-\rho\times avgSTE\right)\right) \quad (2)$$

$$avgSTE = \frac{1}{N}\sum_{n=0}^{N-1} STE(n) \quad (3)$$

$$STE(n) = \sum_{k=n}^{n+m-1} a^2(k) \quad (4)$$

Where N is the total frame number in one-second time window, $STE(n)$ is the short-time energy of frame n, ρ is an experimental parameter so that $\rho\times avgSTE$ acts as the reference to determine if the short-time energy of frame n is low enough to reveal a silence, m is the frame size, $a(k)$ is the signal amplitude at time point k in frame n.

This feature indicates that a frame is regarded as silence if its energy is lower than ρ percent of the average energy in the one-second time window, and ASR is a counter of silence frame in this time window. The lower ASR means fewer silence frames present in music piece, or legato in articulation, and the higher ASR means more silence frames present in music piece, or staccato in articulation.

Figure 2 shows that ASR is a good discriminator of different moods in the facet of music articulation, for example, when $\rho = 0.25$, the values of ASR for "happiness" and "fear" segments are more than 0.15, the values of ASR for "anger" and "sadness" are less than 0.15.

Figure 2. The average silence ratio of four segments of 20s music in different moods: happiness, fear, sadness and anger. The *ASR*s of "happiness" and "fear" segment is relatively high with the value higher than 0.15, the *ASR*s of "sadness" and "anger" segment is relatively low with the value less than 0.15 with $\rho = 0.25$.

5.2. Articulation model

In our scheme, with regard to a piece of music, we use the mean and standard deviation of its *ASR* sequence as two simple features, which are denoted as *mASR* and *vASR*, to model music articulation, we believe that them give us more information about the articulation character of music piece then *ASR* does because they are the global statistics of the feature distribution. Other features describing singing or performing technique such as vibration in music piece can also be considered as options to model articulation, but *mASR* and *vASR* do good job in music mood detection. Figure 5 reveals that "happiness" and "fear" or "sadness" and "anger" music pieces have similar probability distributions on the mean and standard deviation of *ASR*, which implies that these features are useful in modeling music articulation.

6. Music mood detection

To detect music mood, feature space is mapped to mood space. If tempo known, the question raised is how to decide it is fast or slow, the similar question about articulation is how much *ASR* is assigned to staccato or legato. One possible way is to express "fast", "slow", "staccato" and "legato" in fuzzy quantity, in this way, we must decide membership functions for them, which will introduce a significant difficulty to cope with. A simple neural network classifier with three layers, input, hidden, output layer, trained by Back-Propagation algorithm, is employed to detect music mood in this paper, our intention is to express the human perception of temp and articulation in a implied manner. The structure and weight matrix of the neural network provide us with the knowledge about how people regard tempo as being "fast" or "slow" and articulation as being "staccato" or "legato". The input layer of the neural network has three nodes, being *rTEP*, *mASR* and *vASR*, respectively, and the output layer has four nodes representing the four music moods. A little more explanation about the output of the neural network, the output value is a score in [1 0.5 0] with 1, 0.5, 0 representing the piece of music definitely, probably, impossible evoking a specific kind of mood. For example, the score vector of "Happy Birthday to You" is [1, 0, 0, 0] on mood vector [happiness, sadness, anger, fear] because most of us think it evokes "happy" mood.

In training procedure of the neural network, the inputs are features derived from music signal, the outputs are the scores of music piece, they are obtained from musician's estimation, for each piece four scores are given associating with four moods. In detection procedure, the output scores are the estimations about to what degree the corresponding mood is evoked. Under the consideration that composer or performer may probably manipulate multiple moods in a single piece of music, we employ the mood "*score*" instead of explicit judgment such as "yes" or "no" to depict music mood. If $score \geq 0.5$ for a specific output node, the piece of music is in the corresponding mood, if all $score < 0.5$, the representing mood by the output node with largest score value is assigned to the music piece.

7. Visualization and browsing

User interface in the music retrieval system should provide user with convenient tools to search music, generate play list, play music piece, etc. We think that a user interface, which can graphically display the result of retrieval, is more effective than traditional text browsing interface. Under this consideration, we propose to access music database by visualizing it. For example, if a query mood is "happiness", all music pieces in the "happiness" sub-database are displayed on the browser. As has been illustrated in Figure 3, horizontal and vertical coordinates specify *mASR* and *vASR* respectively, the gray level of flat mesh specifies *rTEP*. The browser is capable of zooming in and out to travel in the concerned region. If a user wants to search a piece of music that is happy in mood, fast in tempo and moderate in continuity, he should move his mouse to the low right corner region of browser and click the dark points, the browser zooms in and he can refine his search to locate a music piece.

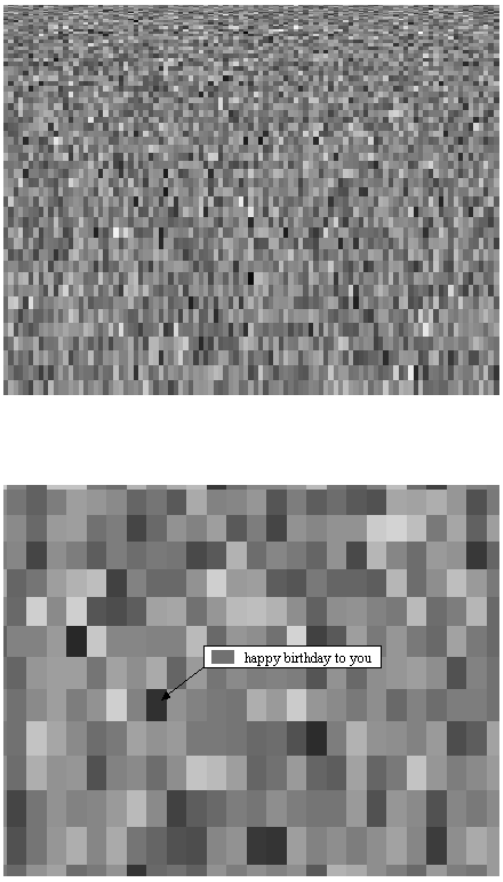

Figure 3. The upper figure is 3D visualization of 100 pieces of "happiness" music; the lower figure zooms in the visualization and "Happy birthday to you" is browsed.

8. Experimental results

We collect 353 pieces of modern popular music containing multiple instruments and vocal singing from Web and personal CD repository, stereo music is converted to single channel signal by averaging the left and right channels, 330 pieces are used as training data, 23 pieces are used as testing data.

8.1. Processing

Each music piece is converted to 22050Hz/mono/16bit raw audio signal; we calculate *rTEP*, and use a frame size of 256 points with 128 points overlapping and a time window of one second to calculate *ASR*, then we calculate the *mASR* and *vASR* for each music piece.

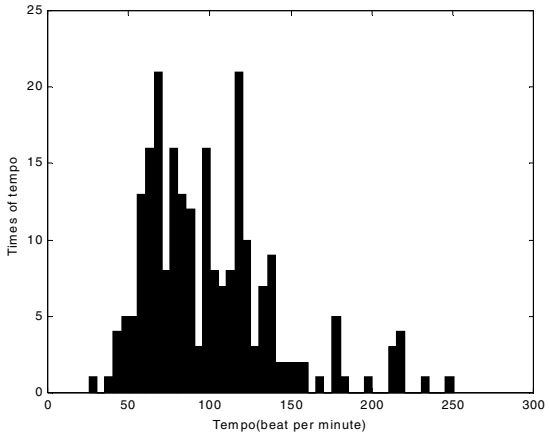

Figure 4. Tempo distribution in the experimental corpus, most tempo is between 50 and 150, the slowest tempo is 28, the fastest tempo is 250.

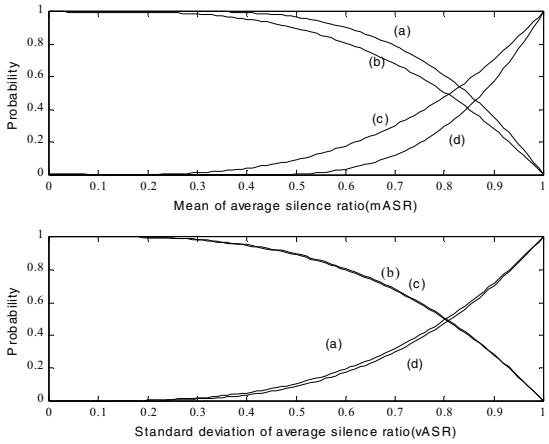

Figure 5. Probability distributions of *mASR* and *vASR*, the curves labeled by (a) (b) (c) (d) represent the probability distributions of "sadness", "anger", "happiness" and "fear" music piece, respectively.

The structure of neural network classifier in our experiment is three input nodes, ten hidden nodes and four output nodes, to train the neural network, [*rTEP* , *mASR* , *vASR*] of music piece is used as the inputs of neural network classifier, the target outputs are scores of music mood. To detect mood of music piece, scores are calculated by feeding the classifier with extracted features from music signal.

8.2. Results

We analysis the tempo distribution (illustrated in Figure 4) of music pieces in our experimental corpus, and derive two parameters in *rTEP* calculation, $s = 20$,

$f = 280$, though they are derived from our test corpus, they are also applicable in larger music database.

We also plot the probability distributions of *mASR* and *vASR* for the experimental corpus (illustrated in Figure 5). For the "happiness" and "fear" music, the probability distributions of mean are similar, and for "sadness" and "anger" music, their probability distributions are also similar, because of their similar articulation manner. We can conclude that music in different moods in differentiable by the mean of average silence ratio. The probability distributions of standard deviation for "happiness" and "fear" or "sadness" and "anger" are not similar, that is, the four labels of music mood are differentiated in articulation, and this evidence supports our computational articulation model.

We perform music mood detection on our experimental corpus, the detection result is given in Table 2.

We calculate the *precision* and *recall* to evaluate the retrieval performance (illustrated in Table 3), the total *precision* is 67% and the total *recall* is 66%.

Table 2. Experimental result on mood detection of 23 pieces of music. H, S, A and F denotes happiness, sadness, anger and fear respectively.

MUSIC PIECE	MOOD	DETECTED MOOD			
		H	S	A	F
1	H	0.84	0.23	0.11	0.41
2	H	0.78	0.34	0.20	0.22
3	S	0.65	0.43	0.32	0.67
4	A	0.09	0.18	0.56	0.26
5	S	0.37	0.45	0.29	0.80
6	H	0.96	0.16	0.29	0.07
7	F	0.46	0.38	0.29	0.76
8	S	0.15	0.67	0.12	0.39
9	A	0.18	0.26	0.34	0.27
10	A	0.49	0.07	0.79	0.03
11	F	0.04	0.29	0.38	0.17
12	S	0.34	0.92	0.41	0.25
13	S	0.26	0.10	0.24	0.09
14	H	0.32	0.01	0.45	0.57
15	A	0.56	0.42	0.89	0.09
16	H	0.81	0.34	0.63	0.13
17	S	0.78	0.23	0.61	0.75
18	H	0.55	0.23	0.54	0.22
19	H	0.35	0.87	0.23	0.32
20	S	0.39	0.72	0.11	0.25
21	S	0.52	0.16	0.30	0.66
22	F	0.09	0.18	0.34	0.90
23	A	0.24	0.31	0.67	0.87

Table 3. Retrieval performance

%	HAPPINESS	SADNESS	ANGER	FEAR
PRECISION	86	75	83	25
RECALL	57	38	100	67

9. Discussion and future research directions

This paper introduces our preliminary work on music retrieval by detecting mood; experimental results show that music mood is computable. Three simple features do a satisfactory job in detecting mood; they also make visualization of music database possible.

Our test corpus is by no means large; if possible, our scheme should be tested on a commonly recognized evaluation corpus to be more convictive..

In this paper, we suppose that tempo and articulation are invariable in music piece, which is not always true, especially for classical music, without this hypothesis, a possible approach to detect mood of music with fluctuant tempo and articulation is, at first, to detect mood in the short segments with invariable tempo and articulation, then derive the mood of the concerned music piece from them by a perceptional solution, we will discuss this approach in future paper.

For our computational articulation model, we only use time domain energy of music signal to calculate articulation feature, we will try other features in future. For the parameter ρ in *ASR*, maybe further analysis should be performed to make clear that if it critical in modeling articulation.

We can also experiment to verify if other cluster algorithms are more powerful than the BP neural network on music mood detection.

10. References

[1] Ghias A., Logan,J. Camberlin,D. and Smith, B. C. "Query by humming: Musical information retrieval in an audio database", *Proc. ACM Int. Conf. On Multimedia*, ACM, San Fancisco, CA, 1995, pp. 231–236.

[2] Feng,Y. Z., Zhuang,Y. T. and Pan,Y. H. "A hierarchical approach: query large music database by acoustic input", *Proc. SIGIR*, July 2002, pp. 441-442.

[3] Kosugi, N., Nishihara, Y., Sakata, T., Yamamuro, M., and Kushima, K. "A practical query-by humming system for a large music database", *Proc of the ACM MM2000*, Marina del Ray, CA, 2000, pp. 333-342.

[4] Feng, Y. Z., Zhuang, Y. T., Pan, Y .H. "Popular music retrieval by independent component analysis", I*SMIR 2002 Conf. Proc.*, October 2002, pp. 281-282.

[5] Liu, M. and Wan, C. "A study of content-based retrieval of mp3 music objects", *Proc. of the Int'l conf on Information and knowledge management*, Atlanta, Georgia, 2001, ACM, pp. 506-511.

[6] Yang, C. "The MACSIS Acoustic indexing framework for music retrieval: an experimental study", *Proc. of third international conference on music information retrieval (ISMIR2002)*, Paris, France, 2002, pp. 53-62.

[7] Cooper, M. and Foote, J. "Automatic Music Summarization via Similarity Analysis", *Proc. of third international conference on music information retrieval (ISMIR2002)*, Paris, France, 2002, pp. 81-85.

[8] Logan, B. and Salomon, A. "A music similarity function based on signal analysis", *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, August 2001.

[9] Rauber, A., Pampalk, E., Merkl, D. "Content-based music indexing and organization", *Proc. SIGIR*, July 2002.

[10] Pickens, J. "A survey of feature selection techniques for music information retrieval", *Proc. SIGIR*, September, 2001.

[11] Rolland, P. Y., Raskinis, G., Ganascia, J. G. "Musical content-based retrieval: an overview of the Melodiscov approach and system", *ACM Multimedia* (1) 1999, pp.81-84.

[12] Goldberg, D., Nichols, D. , Oki, B. and Terry, D., "Using Collaborative Filtering to Weave an Information Tapestry", *Communications of the ACM*, Vol. 35(12), December 1992, pp. 61-70.

[13] Burke, R., "Hybrid Recommender Systems: Survey and Experiments", *User modeling and User-adapted Interaction*, 2002.

[14] Dorai, C. and Venkatesh, S., "Computational Media Aesthetics: Finding meaning beautiful", *IEEE Multimedia*, 8(4), October-December, 2001, pp. 10-12.

[15] Truong, B.T., Venkatesh, S., Dorai, C., "Application of computational media aesthetics methodology to extracting color semantics in film", *ACM Multimedia*, Juan Les Pins, France, 2002.

[16] Adams B., C. Dorai, and S. Venkatesh, "Automated film rhythm extraction for scene analysis", *IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, August, 2001.

[17] Adams B., C. Dorai, and S. Venkatesh, "Towards automatic extraction of expressive elements from motion pictures: Tempo", *IEEE International Conference on Multimedia and Expo*, volume II, New York City, USA, July 2000, pp. 641–645.

[18] Dellaert, F., Polzin, T., Waibel, A. "Recognizing Emotion In Speech", *Proc. ICSLP '96*.

[19] Huber, R., Batliner, A., Buckow, J., Nöth, E., Warnke, V., Niemann, H. "Recognition of Emotion in a Realistic Dialogue Scenario".,*Proc. ICSLP'2000*.

[20] Tzanetakis, G. and Cook, P. "Musical Genre Classification of Audio Signals", *IEEE Transaction on Speech and Audio Processing*, vol.10, July 2002, pp. 293-302.

[21] Farnsworth, P.R., "*The Social Psychology of Music*", Holt, Rinehart and Winston, New York, 1958.

[22] Thayer, R. E. The biopsychology of mood and arousal. New York: Oxford University Press, 1989.

[23] Zillmann, D. "Mood management in the context of selective exposure theory", In M. E. (Ed.), Communication yearbook 23, Thousand Oaks, CA: Sage, pp. 103-123.

[24] Juslin, P.N., "Cue utilization in communication of emotion in music performance: Relating performance to perception", *J. Experimental Psychology*, 26, 2000, pp. 1797-1813.

[25] Dixon, S. "A lightweight multi-agent musical beat tracking system", *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, Melbourne, Australia, 2000, pp. 778-788.

[26] Scheirer, E., "Tempo and beat analysis of acoustic musical signals", *Journal of the Acoustical Society of America*, 103 (1), 1998, pp. 588-601.

[27] Goto, M. and Muraoka, Y. "An audio-based real-time beat tracking system and its applications", *Proc. of the International Computer Music Conference*, Computer Music Association, San Francisco CA, 1998.

[28] Foote, J. "The beat spectrum: a new approach to rhythm analysis", *IEEE International Conference on Multimedia & Expo*, Tokyo, Japan, 2001.