Leveraging Viewer Comments for Mood Classification of Music Video Clips

Takehiro Yamamoto
Kyoto University, Japan
tyamamot@dl.kuis.kyoto-u.ac.jp

Satoshi Nakamura
Meiji University, Japan
JST CREST
satoshi@snakamura.org

ABSTRACT

This short paper proposes a method to classify music video clips uploaded to a video sharing service into music mood categories such as “cheerful,” “wistful,” and “aggressive.” The method leverages viewer comments posted to the music video clips for the music mood classification. It extracts specific features from the comments: (1) adjectives in comments, (2) lengthened words in comments, and (3) comments in chorus sections. Our experimental results classifying 695 video clips into six mood categories showed that our method outperformed the baseline in terms of macro and micro averaged F-measures. In addition, our method outperformed the existing approaches that utilize lyrics and audio signals of songs.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Music Mood Recognition, Music Information Retrieval, User Generated Media

1. INTRODUCTION

Music is a fundamental form of entertainment in our lives, and the expansion of the Web has drastically increased the amount of the digital music contents available for us. Particularly, in Japan, a singing synthesizer software called *Vocaloid* [8] has made it surprisingly much easier for those who have never done so before to create original songs. As a result, such user-generated songs are now uploaded to the Web every day. For example, as of the end of August 2012, over 100,000 music video clips of the user-generated Vocaloid songs (typically with one or more picture images) had been uploaded to *NicoNico Douga*[1], one of the most popular video sharing services in Japan. Furthermore, over 1,000 music video clips of Vocaloid songs are being uploaded to it every month.

While the amout of digital music contents available for us has been rapidly increasing, the ways to search for desired songs are still limited. Usually, users have to retrieve songs on the basis of their meta-data such as artist name, song name, play count, or social tags annotated by users. To offer more ways to search for desired songs, some researchers in the field of Music Information Retrieval (MIR) have recently been working on the music *mood* recognition [7, 9]. If we can offer a mood-based music retrieval system to users, they can search for their desired songs with their subjective information needs: “I feel depressed after quarreling with my friend, so I want to listen a *cheerful* song to clear my mind,” or “I’m unfamiliar with Vocaloid songs, but I love *wistful* songs, so I’d like to listen to popular wistful Vocaloid songs.” In this way, such mood-based retrieval can offer users another way to find songs. In particular, it would be beneficial to novices who know little about the target music domain.

In MIR, most work on the music mood recognition basically relies on the features extracted from the audio signals. Some recently combined the lyrics of the songs with the audio signals [6, 10]. However, mood recognition is much harder than genre or artist recognition [9]. One alternative way for the mood-based retrieval is to use social tags related to the music moods annotated by users. However, according to our preliminary study with 186,987 music video clips on NicoNico Douga, only 5% of them contained the music mood related tags. Similarly, on Last.fm[2], which is a popular music social networking service, only 14% of songs had mood related tags [6]. These results suggest that the simple exploitation of social tags is also insufficient. The goal of our work is to achieve a more effective mood-based music retrieval system to help users search for many songs.

As the first step to this goal, in this paper, we propose a method to classify music video clips uploaded to NicoNico Douga into mood categories by leveraging viewer comments. On NicoNico Douga, viewers can post comments at *arbitrary temporal playback positions* in the video clip while they are watching it (See Section 2). These comments can be seen as viewers’ direct responses to music video clips, thus they should contain useful information to estimate the music mood of the music video clips.

This short paper tackles the following two issues. First, to extract useful features for the mood recognition from viewer comments, we propose a method that combines three feature extraction approaches: (1) adjectives in comments, (2) lengthened words in comments, and (3) comments in chorus sections. In this paper, we examined the effectiveness of the proposed method with a dataset consisting of 695 music video clips in six mood categories (See Section 4.3). Second, to investigate the effectiveness of viewer comments for the music mood recognition, we also compared our

[1] http://nicovideo.jp, http://en.wikipedia.org/wiki/Nico_Nico_Douga

[2] http://www.last.fm

SIGIR’13, July 28–August 1, 2013, Dublin, Ireland.

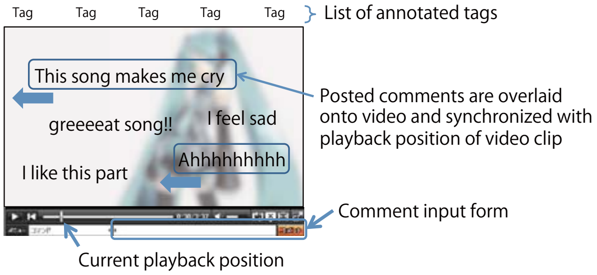

Figure 1: Overview of viewer interface in NicoNico Douga.

method with the existing techniques that use lyric and audio signal features (See Section 4.4).

2. OVERVIEW OF NICONICO DOUGA

NicoNico Douga is one of the most popular video sharing services in Japan. It had about 29.6 million signed-up users as of the end of June 2012, and over 5,000 video clips are uploaded to it every day. Figure 1 shows a viewer interface in NicoNico Douga. One unique feature of this service is that viewers can post comments at arbitrary playback positions to the video clip, and such comments of many other viewers are overlaid directly onto the video and synchronized to a specific playback position. This gives viewers a sense of sharing the viewing experiences of the video with others, who originally watched the video at different times.

In addition to its unique comment feature, like other content sharing services such as YouTube and Flickr, viewers can annotate *tags* to the video. In this work we employ such tags to build the data for the music mood recognition (See Section 4.1).

Recently some researchers have been working on leveraging viewer comments on the video sharing services like YouTube to improve video retrieval [2, 3]. While comment analyses of NicoNico Douga have also attracted the attension of some researchers [13], to our knowledge, our method is the first to use viewer comments for music mood classification.

3. FEATURE EXTRACTION FROM VIEWER COMMENTS

In this work, we treat music mood recognition of a music video clip as a multiclass classification problem. That is, given a music video clip, we classify it into one of the predefined mood categories. Figure 2 shows the overview of our feature extraction methods from viewer comments. To extract effective features from comments, our proposed method combines three feature extraction approaches: (1) adjectives in comments, (2) lengthened words in comments, and (3) comments in chorus sections. The rest of this section describes how each method extracts features from comments.

3.1 Adjectives

One simple approach to obtaining features is to extract Bag-of-Words from viewer comments. However, comments posted to a music video clip may contain various pieces of information, but most have nothing to do with the mood of the song. If we extract unnecessary features, they may deteriorate the classification performance.

In this work, we focused on adjectives, rather than all Bag-of-Words, as features. It is natural to think that the viewers who are listening a wistful song may write comment about the clip such as “I feel *sad*,” or “This song is *tear-jerking*.” In this way, adjectives have strong relationship with music moods.

The method works as follows. The method first splits all the comments into words by using a morphological analyzer. The method then extracts adjectives from the words and uses them as the features for the classification.

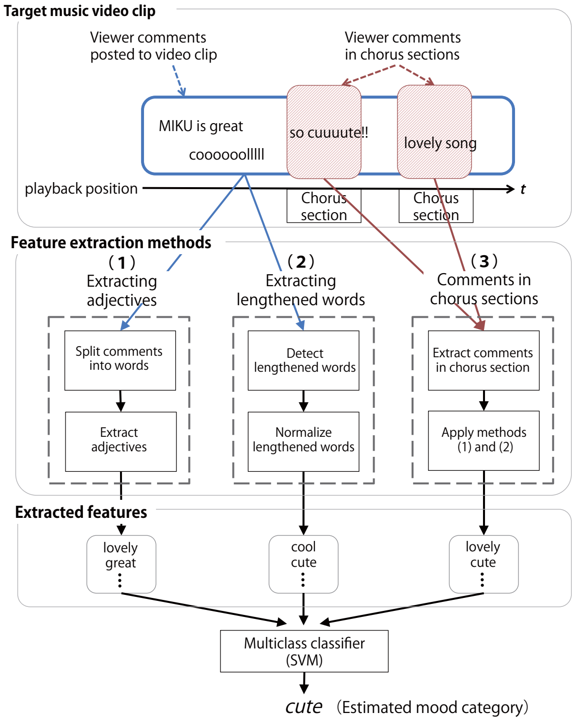
Target music video clip
Viewer comments posted to video clip
Viewer comments in chorus sections
MIKU is great
coooooollllll
so cuuuute!!
lovely song
playback position
t
Chorus section
Chorus section
Feature extraction methods
(1) Extracting adjectives
(2) Extracting lengthened words
(3) Comments in chorus sections
Split comments into words
Extract adjectives
Detect lengthened words
Normalize lengthened words
Extract comments in chorus section
Apply methods (1) and (2)
Extracted features
lovely
great
cool
cute
lovely
cute
Multiclass classifier (SVM)
cute
(Estimated mood category)

Figure 2: Overview of our feature extraction method.

3.2 Lengthened Words

The method described in Section 3.1 applies a morphological analysis to comments and extracts adjectives as features. For viewer comments in NicoNico Douga, however, a morphological analyzer often fails to split the comments into the correct words, since the form of comments are far from standard Japanese. One good English equivalent is tweets on Twitter[3], as reported by Brody and Diakopoulos [1]. Brody and Diakopoulos pointed out that some words in tweets are often *lengthened* by repeating some characters. For example, word “cool” is often lengthened as “cooolll” or “coooooollllllll” on Twitter. Brody and Diakopoulos further pointed out that these lengthened words are strongly related to the sentiments (positive or negative) of the tweets and proposed a method to leverage such lengthened words to analyze these sentiment. On NicoNico Douga, similar to the above case on Twitter, viewers often write comments with such lengthened words. We hypothesized that such lengthened words in comments are also good indicators of the music moods and can compensate for a morphological analysis approach described in Section 3.1.

The method works as follows. The method first detects lengthened words from the comments by using the method proposed by Brody and Diakopoulos [1]. It then obtains the normalized forms of these lengthened words and uses them as features. For example, both “coooollllll” and “cooool” are normalized to “cool” and treated as the same feature.

3.3 Comments in Chorus Sections

Neither approach described in Sections 3.1 and 3.2 considers associated playback positions of viewer comments. However, the likelihood that a comment is related to the mood of the music video clips might depend on its associated playback position. For exam-

[3] http://twitter.com/

Table 1: Dataset used in experiments.

Mood category	# of tags	# of clips
wistful	4	160
cute	9	153
fresh	7	145
cool	9	109
cheerful	5	84
aggressive	4	44
Total	38	695

ple, comments posted before the music actually starts might have nothing to do with the mood of the music. In this paper, we hypothesized that comments posted in the *chorus*[4] section of the music are more likely to relate to the mood of the music, as the chorus sections are a song's most representative and memorable portions.

The method works as follows. The method first detects the chorus sections of the song of the target music video clip. To detect chorus sections of songs, in this work we use the method proposed by Goto [4]. After detecting the chorus sections by using Goto's method, our method then collects comments whose associated playback positions are within the detected chorus sections and extracts the features from them by using the methods described in Sections 3.1 and 3.2.

Note that the three methods described above might extract the same feature in terms of its characters (e.g., all methods might extract "cool" as features). We thus assign different feature IDs to them in order to distinguish them.

4. EVALUATION

4.1 Dataset

To evaluate our method, we built a dataset from music video clips in NicoNico Douga. As for the music moods, many researchers have proposed ways to model moods of music [5, 11]. In this work, we took an approach similar to Hu *et al.*'s work [6]. Hu *et al.* have proposed a method of mood recognition that uses both audio signal and lyric features. To evaluate their method, they created a dataset from the social tags annotated to the songs on Last.fm. They picked up 135 music mood related tags and manually merged them into 18 mood categories. The created mood categories are public to researchers as MIREX [7] Mood Tag Dataset[5]. Their approach, which creates a dataset from social tags, has two merits: (1) flexible mood categories that suit the target domain can be obtained, and (2) ground truth (i.e., which song belongs to which mood category) can be almost automatically created from the relationships between tags and songs.

In our work, we first manually collected 50 tags related to the music moods from NicoNico Douga. Then we carefully merged similar tags into the same group so that the resulting mood groups would be similar to the Mood Tag Dataset. After this operation, we obtained 11 mood categories. Then, for each mood category, we retrieved music video clips that had (1) more than 200 viewer comments and (2) one of the associated tags in the mood category. Finally, we dropped mood categories that had fewer than 40 retrieved video clips, resulting in six mood categories with 695 associated music video clips. Table 1 shows the statistics of the dataset used in the experiments. In the table, "# of tag" represents the associated tags with the mood, and "# of clips" represents the number of music video clips in the mood category.

4.2 Settings

We used Support Vector Machine (SVM) to construct a multiclass classifier. We used LIBSVM[6] to implement the SVM and chose the linear kernel, since the preliminarily experiment showed that the linear kernel outperformed the other kernels. We used MeCab[7] as a Japanese morphological analyzer. Following Hu *et al.*'s work [6], each feature was weighted by the tf-idf weighting.

For each mood category, F-measure was calculated over a five-fold cross validation. We also computed the macro and micro averages over the six mood categories to evaluate the classification performance.

4.3 Experiment #1: Comparison of Feature Extraction Methods from Comments

The purpose of this experiment is to examine the effectiveness of our methods for feature extraction from viewer comments. First, we explain the baseline and proposed methods we used and then discuss their results.

Methods. For the experiment, we prepared one baseline and five proposed methods. For the baseline, BoW uses all the nouns, verbs, and adjectives obtained from viewer comments as features. As for the proposed methods, Adj, Len, and Cho correspond to the methods described in Section 3.1, 3.2, and 3.3, respectively. We also prepared the combination of these methods: Adj+Len uses features extracted by both Adj and Len, and Adj+Len+Cho uses all the features extracted by Adj, Len, and Cho.

Results. Table 2 shows the result of classification performances obtained from the methods described above. First, when we compare the results of Adj, Len, and Cho, we find that Adj achieved the highest macro (.632) and micro (.676) F-measures. In addition, when we compare Adj with BoW, we can see that Adj also achieved higher performance than the baseline in terms of macro and micro F-measures. These results indicate that adjectives in viewer comments play an important role in recognizing the music moods. When we see the results of Len, we found that Len could recognize the "aggressive" category better than the other categories. Note that Len uses lengthened words like "coooolll." Such words might be likely to be posted to express viewers' highly intense emotions, rather than calmness or relaxation. Thus, Len might be useful to recognize the mood categories that are related to highly intense moods (high arousal space in Russel's model [11]).

From the table, we can see that Adj+Len+Cho outperformed the other methods in terms of macro and micro averaged F-measures. This result suggests that combining the different approaches can improve the music mood recognition. In particular, note that Cho considers associated playback positions of viewer comments. This indicates that considering the playback positions of comments can improve the music mood recognition. Although in this work we just combined the features extracted from different methods with the linear kernel, we plan to explore a more sophisticated method to integrate features extracted from these three methods.

4.4 Experiment #2: Comparison of Comments, Lyrics, and Audio Signals

The purpose of this experiment is to examine the effectiveness of our method, which uses viewer comments for the mood recognition, compared with the existing approaches that uses lyric and audio signal features [6].

Methods. Following Hu *et al.*'s work, we prepared two methods named Lyric and Audio, which extract features from the lyrics

[4] a.k.a. refrain

[5] http://music-ir.org/mirex/wiki/2010:Audio_Tag_Classification

[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[7] https://code.google.com/p/mecab/

Table 2: Comparison of F-measure for methods.

		baseline	proposed methods				
		`BoW`	`Adj`	`Len`	`Cho`	`Adj+Len`	`Adj+Len +Cho`
categories	wistful	.636	**.758**	.610	.678	.708	.722
	cute	.627	.731	.708	.727	.780	**.782**
	fresh	549	**.691**	.492	.568	.621	.638
	cool	.638	.659	.635	.577	**.771**	.672
	cheerful	.228	.480	.246	.334	.432	.432
	aggressive	**.792**	.469	.739	.528	.744	.742
Macro average		.578	.632	.572	.568	.659	**.665**
Micro average		.586	.676	.578	.607	.674	**.683**

and audio signals, respectively. For the lyric features, we collected the lyrics for the music video clips in the dataset from the lyric database[8]. We then split the lyrics into words and extracted nouns, verbs, and adjectives from them for the lyric features, weighting with the tf-idf weighting. As for the audio signal features, also following Hu *et al.*'s work, we used MARSYAS [12], the best performing audio system evaluated in the MIREX 2007 audio mood classification task [7]. MARSYAS extracts spectral features such as Spectral Centroid, Rolloff, and MFCCs. We extracted the audio signals from the music video clips in the dataset and then applied MARSYAS to them to extract 63 dimensional audio features.

In this experiment, we compare five methods. `Comment` represents the method that extracts features from viewer comments and exactly equals to `Adj+Len+Cho` described in Section 4.3. `Lyric +Audio` use both features extracted by `Lyric` and `Audio`, and `Comment+Lyric+Audio` uses all features extracted from `Comment`, `Lyric`, and `Audio`.

Results. Table 3 shows the classification results of each method. From the table, when we compare the results of `Comment`, `Lyric`, and `Audio`, we can see that `Comment` outperformed the others. This suggests that the viewer comments are useful resources for the mood recognition. Hu *et al.* reported that the classification accuracy is slightly improved when combining lyric and audio signal features are combined. The similar results can be seen from the results of `Lyric`, `Audio`, and `Lyric+Audio`: the classification performance of `Lyric+Audio` improved on these of `Lyric` and `Audio`.

Finally, we can see that `Comment+Lyric+Audio` outperformed the other methods in terms of the macro (.670) and micro (.693) averaged F-measures. Although our dataset is not large, this result indicates that combining features extracted from different sources can enhance the performance of the music mood recognition.

One primary limitation of our method is that it cannot be applied to music video clips that have few comments, as they have just been uploaded a few days previously. In contrast, a classification method based on audio signals can be applied to such new videos. To achieve a more flexible classification method, we plan to integrate comments, lyrics, and audio signals differently, rather than just combing features from them. For example, for new video clips, we first apply an audio-based classification method, and a few days later when the video clips receive enough comments, a comment-based classification will be applied to it. We believe such a flexible integration of different methods might be an interesting research topic for music mood classification.

Table 3: Comparison of F-measure for methods.

		`Comment`	`Lyric`	`Audio`	`Lyric +Audio`	`Comment +Lyric +Audio`
categories	wistful	.722	.490	.470	.561	**.736**
	cute	**.782**	.498	.505	.518	.772
	fresh	.638	.473	.311	.457	**.678**
	cool	**.672**	.424	.333	.466	.663
	cheerful	.432	.203	.072	.227	**.443**
	aggressive	**.742**	.080	.119	.080	.723
Macro average		.665	.362	.302	.385	**.670**
Micro average		.683	.434	.379	.465	**.693**

5. CONCLUSIONS

In this study, we examined the effectiveness of viewer comments for the music mood classification. In the future, we plan to continue this research and develop a more sophisticated method for extracting features from viewer comments. Also, we would like to explore a method to integrate comments, lyrics, and audio signals for more flexible and reliable music mood recognition.

6. ACKNOWLEDGMENTS

This work was supported in part by CREST, JST and KAKENHI (#23680006, #24240013).

7. REFERENCES

[1] S. Brody and N. Diakopoulos. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proc. of EMNLP*, pages 562–570, 2011.

[2] C. Eickhoff, W. Li, and A. de Vries. Exploiting user comments for audio-visual content indexing and retrieval. In *Proc. of ECIR*, to appear, 2013.

[3] K. Filippova and K. Hall. Improved video categorization from text metadata and user comments. In *Proc. of SIGIR*, pages 835–842, 2011.

[4] M. Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5), pages 1783–1794, 2006.

[5] K. Hevner. Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48(2), pages 246–268, 1936.

[6] X. Hu, J. Downie, and A. Ehmann. Lyric text mining in music mood classification. In *Proc. of ISMIR*, pages 411–416, 2009.

[7] X. Hu, J. Downie, C. Laurier, M. Bay, and A. Ehmann. The 2007 MIREX audio mood classification task: Lessons learned. In *Proc. of ISMIR*, pages 462–467, 2008.

[8] H. Kenmochi and H. Ohshita. Vocaloid–commercial singing synthesizer based on sample concatenation. In *Proc. of INTERSPEECH*, pages 4009–4010, 2007.

[9] Y. Kim, E. Schmidt, R. Migneco, B. Morton, P. Richardson, J. Scott, J. Speck, and D. Turnbull. Music emotion recognition: A state of the art review. In *Proc. of ISMIR*, pages 255–266, 2010.

[10] C. Laurier, J. Grivolla, and P. Herrera. Multimodal music mood classification using audio and lyrics. In *Proc. of ICMLA*, pages 688–693, 2008.

[11] J. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), pages 1161–1178, 1980.

[12] G. Tzanetakis and P. Cook. MARSYAS: A framework for audio analysis. *Organised sound*, 4(3), pages 169–175, 1999.

[13] K. Yoshii and M. Goto. MusicCommentator: Generating comments synchronized with musical audio signals by a joint probabilistic model of acoustic and textual features. In *Proc. of EC*, pages 85–97, 2009.

[8] http://www5.atwiki.jp/hmiku/