Can Social Comments Contribute to Estimate Impression of Music Video Clips?

Shunki Tsuchiya[1], Naoki Ono[1], Satoshi Nakamura[1],
and Takehiro Yamamoto[2]

[1] Meiji University, 4-21-1 Nakano, Nakano-ku, Tokyo, Japan
[2] Kyoto University, Yoshida Hommachi, Sakyo-ku Kyoto-shi, Kyoto, Japan
bad.ukr.mbr.pr@gmail.com

Abstract. The main objective of this paper is to estimate the impressions of music video clips using social comments to achieve impression-based music video clip searches or recommendation systems. To accomplish the objective, we generated a dataset that consisted of music video clips with evaluation scores on individual media and impression types. We then evaluated the precision with which each media and impression type were estimated by analyzing social comments. We also considered the possibility and limitations of using social comments to estimate impressions of content. As a result, we revealed that it is better to use proper parts-of-speech in social comments depending on each media/impression type.

Keywords: Estimating Impression, Music Video Clip, Social Comments.

1 Introduction

Due to the spread of consumer-generated media (CGM) websites such as *YouTube* and *Nico Nico Douga*, and the advancement of DTM software such as *VOCALOID* [14], the number of *music video clips*, which are composed of music and a video, on the Web has dramatically increased. A standard method of searching for these music video clips is to input information such as an artists' names, song titles, and tags provided. This search methods makes it possible to find the target music video clip directly.

However, as this method requires users to know information on music video clips in advance, it sometimes is not easy to find the target clips. To solve this, researchers in the field of music searches have been actively researching ambiguous searches based on the user's subjective impressions such as cheerful or sorrowful to solve such problems. If searches based on impressions become possible, the users will be able to search from a new viewpoint. In addition, we can expect users to be able to find new music video clips.

To realize the impression-based music video clip search, we have to evaluate and provide subjective impressions on individual music video clips in advance. However, as previously explained, since the number of music video clips has been increasing explosively, it is too difficult for us to evaluate the impressions of all music video clips. Thus, we need to mechanically estimate the impressions of music video clips. Nevertheless, it is not easy to mechanically estimate the impressions of music video clips that viewers would have because music video clip consists of only music and video.

To achieve this, we decided to use comments written on music video clips on some website. For example, users can freely post comments in order to show appreciation for authors, to communicate with others, to express their feelings, to add explanations and lyrics and so on while viewing a music video clip on *Nico Nico Douga* in Japan and *BiliBili Douga* in China. We regard these comments as the viewers' subjective impressions for the music video clips, and make use of them for mechanical estimation.

Although we conducted the impression estimation of music video clips using comments in our past work [4], we only looked at adjectives in comments, and we did not consider other parts-of-speech. However, we thought that along with adjectives, other parts-of-speech can also be an essential factor to estimate impressions. Thus, we examine what parts-of-speech in comments should be considered for impression estimation.

Also, the past research [4] used the whole of a music video clip for the estimation. However, the most exciting part of the structure of music is known to be a chorus part [13]. Therefore, we assume that the chorus part decides the impression that the viewers would receive, and decided to estimate the impressions of the chorus part of the music video clip only.

A music video clip normally consists of music and a video. As a result, people may focus on different *media types* (i.e., music, video, or combined) of the music video clip when making a search for it based on the impression. For example, one may search for music video clips of *happy* songs, while others may search for those of *cool* video picture. In addition, different people may post comments on different *media types* of a music video clip. For example, one may post comments for music video clips to express "the songs are *happy*", while the others may post comments to express "*cool* video picture." Thus, we focused on social comments on *Nico Nico Douga* and examined the possibility of estimating the impressions of music video clips using comments. At that time, we also considered media types that are music only, video picture only, and combined.

In this paper, we generated the impression evaluation dataset which is an evaluation of eight different types of impressions for each of the three media types (music only, video picture only, and both) for the chorus part of 500 music video clips. In addition, we collected social comments on the chorus part of those music video clips and generated 12 types of bag-of-words based on a particular part-of-speech used in comments. Then, we tested these bag-of-words of estimating the impressions of music video clips. In addition, we examined the accuracy of the estimation with which impressions were estimated by support vector machines (SVMs) using these bag-of-words.

The main contributions of this paper are below.

- We generated the impression dataset of chorus part of 500 music video clips in three media types (music only, video picture only, and combined).

- We revealed that it is better to use proper parts-of-speech in social comments depending on each media/impression type.

2 Related Work

There have been various kinds of researches on estimating impressions of contents of music video clips.

Some researchers initially made estimates of impressions of songs [10, 11]. These researchers improved the accuracy of estimates with not only acoustic features but also subjective features like lyrics. They also estimated the subjects' impressions of videoss [12]. This research disclosed estimates with high levels of accuracy using not only video features but also subjective features such as viewer's expressions.

There have also been many researches on estimates of impressions of music video clips that we have been targeting [8, 9]. The researchers focused on the fact that music video clips combine music and videos, and estimated impressions by combining these characteristics. As a result, although it is possible to estimate impressions with high levels of accuracy, features of music and images are machine enemy features, where no human emotions are reflected. Therefore, we considered that better estimates would be possible using subjective features like those in the researches explained above [10–12].

Therefore, there is a research that has focused on comments provided to music video clips as one of the subjective characteristics of these clips [5, 6]. These researchers have estimated impressions using comments posted on YouTube. However, since comments unrelated to the movies such as conversations between users are posted, we cannot use many of them to estimate impressions. Here, Nico Nico Douga, which is the most popular CGM website in Japan, has a function to provide comments in real time to the video. These comments can be considered to express impressions that users directly felt in real time. In fact, there actually is a research that estimated the impressions of music video clips using these comments [4]. Our research treated adjectives in the comments and the length of the comments as comment features. We focused on the parts-of-speech in the comments and analyzed the accuracy with which impressions were estimated.

There are also various approaches to the impression class. First, there is a research on impressions of clustering of songs [1]. This research clustered the impressions of music into eight groups. Russell also proposed a valence-arousal space as a model of estimating impressions of music [2]. Valence involves pleasure-discomfort, and arousal is a dimension expressing arousal-sedation, which is the idea of expressing an impression in these two dimensions. In our study, we estimated and analyzed the impression of valence-arousal space, the impression of music information retrieval evaluation exchange (MIREX), and the impression of "cute" which is frequently used in Nico Nico Douga.

3 Generating the Impression Evaluation Dataset

In this paper, we generate the impression evaluation dataset of music video clips. The dataset covers the chorus part of music video clips. This dataset also divides one music video clip into three media types (music only, video picture only, and music video clip (combined)), and three or more subjects evaluated eight impressions for each.

We collected target 500 music video clips from March 26, 2015, to June 18, 2016. The music video clips to be evaluated were tagged "VOCALOID" from videos posted on *Nico Nico Douga* and had the large number of views. In addition, we extracted 30 seconds of the music video clip from 5 seconds before the start of the chorus part estimated by refrain detection (RefraiD) [13]. The reason why we chose to extract the clip 5 seconds before the timing detected a chord is that the change from pre-chorus to the chorus would be also important. Most of the music videos targeted this time were those with chorus part less than 25 seconds. In addition, we watched and checked 500 music video clips, but there was no case where chorus part was detected incorrectly.

The eight impressions were composed of five impressions used in MIREX [3], which is a music information search workshop, two impressions called valence-arousal space proposed by Russell et al. [2], and one impression called "cute" used in the research of Yamamoto et al. [4]. Table 1 summarizes the eight impressions used in the dataset. The "impression names" in the table are labels representing the impressions that have been given for convenience. In addition, "adjectives representing impression" express the impression classes when collecting the evaluation value from subjects in dataset construction.

Table 1. 8 Impressions in dataset

Impression names	Adjectives representing impressions
C1（exciting）	Exciting, bustling, proudly, & dignified
C2（cheerful）	Cheerful, happy, hilarious, & comfortable
C3（painful）	Painful, gloomy, bittersweet, & sorrowful
C4（fierce）	Fierce, aggressive, emotional, & active
c5（humorous）	Humorous, funny, strange, & capricious
C6（cute）	Cute, lovely, awesome, tiny, &
Valence	Bright feelings & fun Dark feelings, sad,
Arousal	Fierce, aggressive, & bullish Gentle, passive, & bearish

For the evaluation, we presented one of the media for 30 seconds to subjects.After watching it, they answered each impression with a five rank Likert scale. The impression evaluation dataset was evaluated on a five rank Likert scale from one (strongly disagree) to five (strongly agree) for C1 to C6, -2 (dark feelings and sad) to +2 (bright feelings and fun) for valence and -2 (gentle, passive, and bearish) to +2 (fierce, aggressive and bullish) for arousal. When they finished answering, the next content is presented. We present at random regardless of media type. We asked subjects to evaluate using the Web interface in the above procedure.

To make it easier to compare C1 to C6 and valence-arousal, they were converted to -2 to +2 by decreasing the evaluation values of one to five to -3. After that, we calculated the average of three subjects for the impression evaluation value and used it as the evaluation value for each media and impression type in this paper.

We published this dataset at http://nkmr.io/mood/.

4 Evaluation Experiment

We conducted an evaluation experiment using the impression evaluation dataset to investigate whether evaluations made by people for the impressions of music video clips could be mechanically estimated using social comments.

We tested and verified in the evaluation experiment by using SVMs whether impressions having an evaluation of more than a certain value could be mechanically estimated in the impression evaluation dataset. Two sets of music video clips (high and low evaluation groups) were specially constructed based on the impression evaluation value for each media/impression type. In addition, we divided each dataset into learning and test data. We evaluated the efficiency of classification using the high evaluation group from social comments by learning and testing it with SVMs and performing cross-validation.

First of all, we will describe methods of collecting social comments and generating bag-of-words to perform SVMs, and further I will explain the basic evaluation to consider the amount of data. In addition, each method of generating bag-of-words indicated how much could be estimated by each media/impression type. Based on the results, we will discuss the appropriate method of bag-of-words generation to estimate impressions in each media/impression type.

4.1 Generation of Bag-of-Words for Music Video Clips

We gathered comments given to the music video clips corresponding to the impression evaluation dataset to consider the accuracy with which each media/impression type of a music video clip was estimated from social comments. We specifically collected all comments on the relevant music video clips using the Nico Nico Douga application programming interface (API) on July 23, 2015, and gathered 860,455 comments. Comments posted to the chorus part based on the start and end times of each music video clip were extracted after that. We extracted 132,036 comments (264.1 on average per music video clip) by doing this processing.

We next generated a bag-of-words for music video clips from the social comments. We first morphologically analyzed comments on the chorus part of each extracted music video clip using MeCab [15] and divided them into words. After that, the number of occurrences of each word was taken as a bag-of-words for the music video clips.

We prepared 12 kinds of methods depending on the parts-of-speech used for a bag-of-words generation for the research discussed in this paper.

The first method involved all parts-of-speech. The second method involved four parts-of-speech. Adjectives were considered to show impressions, nouns and verbs were thought to have features presented by the music video clips, and adverbs were considered to express the degree of impression, such as “more” or “very.” We also prepared a method that combined two parts-of-speech and a method that used all four parts-of-speech. Table 2 summarizes all of these method names and the parts-of-speech we used.

Table 2. Methods of bag-of-words generation

Method names	Parts-of-speech used
All method	All parts-of-speech
All2 method	Nouns, Verbs, Adjectives, Adverbs
Noun method	Nouns
Verb method	Verbs
Adj method	Adjectives
Adv method	Adverbs
Noun-verb method	Nouns, Verbs
Noun-adj method	Nouns, Adjectives
Noun-adv method	Nouns, Adverbs
Verb-adj method	Verbs, Adjectives
Verb-adv method	Verbs, Adverbs
Adj-adv method	Adjectives, Adverbs

4.2 Basic Evaluation of Impression Classification

As described in the previous subsection, two sets of music video clips (high and low evaluation groups) were constructed based on the impression evaluation value, and we determined whether the machine could judge the music animation of the high evaluation group for each media/impression type. More specifically, music video clips having an evaluation value of greater than or equal to one were first set as a high evaluation group, and those having minus one or less were set as a low evaluation group to construct a music video clips set. We next divided each music video clip set into five groups and performed five-fold cross-validation using four of them as training data and the other as test data, and calculated the precision of the high evaluation group.

We first evaluated fundamentals in machine learning. Tables 3 and 4 summarize the number of music video clips for each media/impression type of the constructed high and low evaluation groups. “Movie” means music video clips, “Audio only” means music, and “Visual only” means videos. Also, “V” means Valence and “A” means Arousal in the tables below.

Machine learning was performed based on these sets of music video clips by using each bag-of-words. However, a problem with imbalanced data occurred probably because there was bias in the number of music video clips depending on the media/impression type (the number of Audio-C3 and Visual-C1was small.) After this, we under-sampled each media/impression type, and made the number of music video clips the same in an experiment and evaluated them.

Table 3. No. of music video clips in high evaluation group

	C1	C2	C3	C4	C5	C6	V	A
Movie	76	105	87	54	83	104	101	150
Audio	133	127	46	69	49	73	124	178
Visual	21	50	142	49	81	78	57	111

Table 4. No. of music video clips in low evaluation group

	C1	C2	C3	C4	C5	C6	V	A
Movie	105	169	191	209	178	215	62	94
Audio	65	92	232	195	180	209	61	43
Visual	252	272	165	247	207	234	96	155

4.3 Results

Tables 5 to 16 summarize the average precision for the high evaluation group using the SVMs of each media/impression type when we generated the bag-of-words with all the preparation methods. An experiment was also carried out. In addition, each table shows a value of 0.8 or more in pink and a value of 0.6 or less in blue.

Table 5. Precision of All methods

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.720	0.830	0.713	0.765	0.718	0.758	0.783	0.777	0.758
Audio	0.742	0.671	0.612	0.661	0.600	0.712	0.704	0.744	0.681
Visual	0.611	0.680	0.752	0.714	0.603	0.797	0.660	0.743	0.695
Average	0.691	0.727	0.692	0.713	0.640	0.756	0.712	0.755	0.711

Table 6. Precision of All2 method

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.645	0.814	0.705	0.765	0.728	0.792	0.694	0.822	0.745
Audio	0.738	0.658	0.566	0.750	0.725	0.787	0.736	0.778	0.717
Visual	0.880	0.786	0.390	0.725	0.564	0.776	0.814	0.870	0.725
Average	0.754	0.753	0.554	0.747	0.672	0.785	0.748	0.823	0.730

Table 7. Precision of Noun method

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.575	0.720	0.644	0.653	0.704	0.680	0.646	0.652	0.659
Audio	0.698	0.606	0.528	0.621	0.721	0.661	0.708	0.650	0.649
Visual	0.700	0.640	0.608	0.600	0.620	0.688	0.552	0.641	0.631
Average	0.658	0.655	0.593	0.625	0.682	0.676	0.635	0.648	0.647

Table 8. Precision of Verb method

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.667	0.627	0.440	0.544	0.642	0.714	0.575	0.574	0.597
Audio	0.615	0.622	0.133	0.658	0.587	0.500	0.600	0.551	0.533
Visual	0.588	0.549	0.606	0.517	0.584	0.573	0.508	0.654	0.572
Average	0.623	0.599	0.393	0.573	0.604	0.596	0.561	0.593	0.568

Table 9. Precision of Adj method

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.733	0.869	0.710	0.750	0.667	0.838	0.650	0.842	0.757
Audio	0.667	0.635	0.595	0.667	0.581	0.775	0.706	0.733	0.669
Visual	0.714	0.736	0.733	0.759	0.536	0.829	0.603	0.850	0.720
Average	0.705	0.747	0.679	0.725	0.595	0.814	0.653	0.808	0.716

Table 10. Precision of Adv method

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.618	0.586	0.522	0.576	0.520	0.481	0.556	0.603	0.557
Audio	0.679	0.600	0.580	0.537	0.545	0.481	0.642	0.538	0.575
Visual	0.879	0.759	0.211	0.632	0.519	0.451	0.777	0.805	0.629
Average	0.725	0.648	0.438	0.582	0.528	0.471	0.658	0.649	0.587

Table 11. Precision of Noun-verb method

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.687	0.699	0.648	0.620	0.681	0.714	0.661	0.636	0.668
Audio	0.683	0.580	0.489	0.642	0.689	0.672	0.729	0.658	0.642
Visual	0.881	0.760	0.308	0.614	0.595	0.639	0.805	0.859	0.682
Average	0.750	0.680	0.482	0.625	0.655	0.675	0.732	0.718	0.665

Table 12. Precision of Noun-adj method

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.662	0.854	0.690	0.780	0.750	0.778	0.694	0.800	0.751
Audio	0.754	0.644	0.612	0.750	0.707	0.772	0.740	0.806	0.723
Visual	0.888	0.792	0.409	0.706	0.657	0.768	0.821	0.874	0.739
Average	0.768	0.763	0.570	0.745	0.705	0.773	0.752	0.827	0.738

Table 13. Precision of Noun-adv method

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.592	0.714	0.644	0.654	0.722	0.673	0.656	0.649	0.663
Audio	0.672	0.589	0.538	0.621	0.711	0.661	0.694	0.632	0.639
Visual	0.879	0.763	0.372	0.636	0.622	0.683	0.805	0.852	0.701
Average	0.714	0.689	0.518	0.637	0.685	0.672	0.718	0.711	0.668

Table 14. Precision of Verb-adj method

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.781	0.811	0.711	0.684	0.667	0.856	0.652	0.784	0.743
Audio	0.692	0.627	0.520	0.714	0.682	0.740	0.673	0.707	0.669
Visual	0.921	0.734	0.400	0.734	0.511	0.764	0.779	0.871	0.714
Average	0.798	0.724	0.544	0.711	0.62	0.787	0.701	0.787	0.709

Table 15. Precision of Verb-adv method

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.667	0.568	0.535	0.531	0.657	0.630	0.600	0.660	0.606
Audio	0.677	0.560	0.458	0.566	0.587	0.513	0.589	0.581	0.566
Visual	0.882	0.729	0.250	0.622	0.488	0.529	0.724	0.814	0.629
Average	0.742	0.619	0.414	0.573	0.577	0.557	0.638	0.685	0.601

Table 16. Precision of Adj-Adv method

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.700	0.837	0.679	0.690	0.681	0.848	0.695	0.844	0.746
Audio	0.733	0.646	0.581	0.634	0.683	0.743	0.667	0.718	0.675
Visual	0.911	0.765	0.477	0.653	0.622	0.757	0.840	0.884	0.738
Average	0.781	0.749	0.579	0.659	0.662	0.783	0.734	0.815	0.720

First, we found that the value of the All2 method was more significant than that of the All method by more than 0.8 for each media/impression type, when comparing them, and the overall average value was also high. However, the value of C3 (painful) impression was low in all media types.

Next, by comparing methods using only one part-of-speech, we can see that the Noun, Verb, and Adv methods were not as highly accurate in estimation as the Adj method. Although the Adv method had high values that slightly exceeded 0.8, low values below 0.6 were often found. However, the Adj method had many values that exceeded 0.8, and it particularly demonstrated that the precision of C6 (cute) and Arousal was high.

High values increased by combining parts-of-speech for the method using two parts-of-speech; it especially indicated that the method achieved many high values including those for adjectives. Furthermore, high values that exceeded 0.8 for Audio were only found for the Arousal of the Noun-Adj method out of all the methods. However, C3 achieved no high values for any of the methods, and we found that there were many low values below 0.6.

Visual-C1, Movie-C2, and Visual-Arousal attained relatively high values for each media/impression type, regardless of which method was used. We can also see that there is some bias in the media/impression type with high values.

5 Discussion

We found that the accuracy of estimation using social comments differed depending on each method and each media/impression type.

The All2 method achieved much higher values than the All method, and the overall average value was higher for the All2 method. This might be because all types of written expressions including symbols such as parentheses and emoticons, which are hardly thought to represent impressions, were used in the All method. However, we found that the accuracy of C3 for the All method was higher than that of the All2 method in all media types. The Visual-C3 of the All2 method was 0.39, which is especially low. We considered from this that the parts-of-speech excluded by the All2 method were factors in improving the accuracy of estimating C3.

A high value appears in the Adj method using only one part-of-speech; however, we can see that the other three methods do not have high values and the overall average value is also very low. From this, we considered that nouns, verbs, and adverbs were not used much to express impressions, or words used for impressions did not have features. Therefore, we considered that users most often expressed impressions using adjectives and that the words that were used had features.

Next, values exceeding 0.8 increased for each media/impression type in a method that combined two parts-of-speech, unlike a method using only one part-of-speech. Therefore, we considered that the method that used two parts-of-speech was useful. In particular, the results obtained for Visual-valence of Noun-verb and Noun-adv methods were high; however, the results for the Visual-valence of Noun, Verb, and Adv methods were low. We could see from this that the combination of parts-of-speech improved the accuracy of estimation. Since the results differed depending on the combination of parts-of-speech used in bag-of-words generation, it can be assumed that the accuracy of estimation will be higher by combining parts-of-speech not used in this research with other parts-of-speech. However, the value for C3 was lower in all combinatory methods because the parts-of-speech used in this experiment made it difficult to reveal features, and we expect to improve this using the parts-of-speech. However, nouns, verbs, adjectives, and adverbs are major components to construct sentences, and it is difficult to estimate that C3 is lower from the social comments using these parts-of-speech. Moreover, we obtained high values for C6 (cute) and Arousal for the impression type. Therefore, these impressions were considered to be easy to estimate from social comments. The main reason for the higher values was considered to be because the words used in the high evaluation group had features.

For example, users often expressed the impression of C6 (cute) with the word “cute.” Hence, we thought that C6 was able to learn well due to the features. There were also media/impression types that had relatively high values with all methods, such as Visual-C1 (proudly), Movie-C2 (cheerful), and Visual-arousal. We expect that these media/impression types will be easy to estimate from social comments. Therefore, by analyzing what features (number of comments and words used) were used in the comments, we aim to improve the accuracy of impressions in other media/impression types.

Table 17. Method that yielded highest values in each media/impression type

	C1	C2	C3	C4	C5	C6	V	A
Movie	Verb-adj	Adj	All	Noun-adj	Noun-adj	Verb-adj	All	Adj-adv
Audio	Noun-adj	All	Noun-adj	Noun-adj	All2	All2	Noun-adj	Noun-adj
Visual	Verb-adj	Noun-adj	All	Adj	Noun-adj	Adj	Adj-adv	Adj-adv

Table 18. Highest value for each media/impression type

	C1	C2	C3	C4	C5	C6	V	A	Average
Movie	0.781	0.869	0.713	0.780	0.750	0.856	0.783	0.844	0.797
Audio	0.754	0.671	0.612	0.750	0.725	0.787	0.740	0.806	0.731
Visual	0.921	0.792	0.752	0.759	0.657	0.829	0.840	0.884	0.804
Average	0.819	0.777	0.692	0.763	0.711	0.824	0.788	0.845	0.777

Tables 17 and 18 lists the methods and the values that yielded the highest value in each media/impression type. Table 17 indicates that methods that included adjectives had the highest values with all the media/impression types. We considered from this that people use adjectives when expressing impressions, and features are likely to appear in the adjectives. We also found that adjectives are an important parts-of-speech when estimating impressions of music video clips from social comments.

Table 18 indicates that values that exceed 0.75 appear in 20/24 media/impression types (three media × eight impressions). Since the evaluation value of the dataset used in this paper was the averaged value of the evaluations by three people, there is a blur in the evaluation value. Therefore, we considered that accuracy that exceeded 0.75 was relatively effective. In particular, values that exceeded 0.8 could be classified with accuracy that was as high as 80%, so we considered those to be an effective value.

There was a clear difference when we compared the average of Audio and Visual types. Users tended to comment on the video from this, and we considered that estimates from the comments were useful concerning the impressions of the video. We only used VOCALOID songs in this paper. Therefore, there is a possibility that comments on characters will be made regardless of the music when a character such as Hatsune Miku appears in a video. We plan to analyze this carefully.

We considered that estimating the impressions of music video clips from social comments could be done by separately using methods that were suitable for each media/impression type, based on the results above when estimating the impressions of music video clips. Moreover, if it is possible to estimate the impressions of all media types, we expect that highly accurate estimates of impressions of music video clips will be possible by combining them with researches on combining the impressions of music and video.

6 Conclusion

In this paper, we generated the impression evaluation dataset that consisted of 500 music video clips, three media, and eight impressions, and analyzed the possibility of estimating impressions for each media/impression type from social comments using this dataset. We created bag-of-words for music video clips and obtained the results from estimating impressions using SVMs for each media/impression type and discussed their usefulness. When generating bag-of-words, we mainly used four parts-of-speech and their combinations, compared each method, and found effective methods for each media/impression type. As a result, we found that there was a difference in the accuracy of estimation of each method and that methods that included adjectives yielded the highest values in all media/impression types.

When estimating the impressions of music video clips, from the results in this research, we considered that estimates of impressions was possible using the most effective method for each media/impression type. Therefore, social comments can contribute to estimate impression of music video clips. However, the highest values for Audio-C2 (hilarious), Audio-C3 (painful), and Visual-C5 (humorous) were not high as each of them were 0.671 for Audio-C2, 0.612 for Audio-C3, and 0.657 for Visual-C5. We aim to improve accuracy in this regard by not only estimating impressions from social comments, but also estimating impressions in combination with other features such as sound and video. This was also considered to be similar not only to media/impression types, which had low values, but also to all of them.

We evaluated the accuracy of estimation using classification accuracy in this research; however, we considered that searches based on higher accuracy in impressions will become possible by concretely estimating the evaluation value. Therefore, we intend to explore specific methods of estimating the evaluation value in the future. In addition, we considered that there was blurring in the evaluation value in the impression evaluation dataset used in this research because there were only three evaluators. Furthermore, since we did not investigate the influence of the number of comments or what kind they were, we plan to investigate their impact with Yamamoto and Nakamura [4] in the future.

Acknowledgments. This work was supported in part by JST ACCEL Grant Number JPMJAC1602, Japan.

References

1. Hevner, K.: Experimental studies of the elements of expression in music. The American Jounal of Psychology 48(2), 246–268 (1936).
2. Russell, James A.: A Circumplex Model of Affect. Journal of Personality and Social Psychology 39(6), 1161–1178 (1980).
3. Xiao Hu, J. Stephen Downie, Cyril Laurier, Mert Bay, Andreas F. Ehmann.: The 2007 MIREX audio mood classification task: Lessons learned. In: 9th International Conference on Music Information Retrieval, pp. 14–18. ISMIR, Philadelphia (2008).
4. T. Yamamoto, S. Nakamura.: Leveraging Viewer Comments for Mood Classification of Music Video Clip. In: 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 797–800. SIGIR, Dublin (2013).
5. C. Eickhoff, W. Li, A. de Vries.: The Exploiting user comments for audio-visual content indexing and retrieval. In: 35th European Conference on Advances in Information Retrieval, pp. 38–49. ECIR, Moscow (2013).
6. K. Filippova, K. Hall.: Improved video categorization from text metadata and user comments. In: 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 835–842. SIGIR, Beijing (2011).
7. K. Yoshii, M. Goto.: MusicCommentator: Generating Comments Synchronized with Musical Audio Signals by a Joint Probabilistic Model of Acoustic and Textual Features. In: 8th International Conference on Entertainment Computing, pp. 85–97. ICEC, Paris (2009).
8. E. Acar, F. Hopfgartner, S. Albayrak.: Understanding Affective Content of Music Video through Learned Representations. In: 20th Anniversary International Conference on MultiMedia Modeling, pp.303–314. MMM, Dublin (2014).
9. Y. Ashkan, S. Evangelos, F. Nikolaos, E. Touradj.: Multimedia Content Analysis for Emotional Characterization of Music Video Clips. EURASIP Journal on Image and Video Processing, 1–10 (2013)
10. X. Hu, J. Downie, A. Ehmann.: Lyric Text Mining in Music Mood Classification. In: 10th International Society for Music Information Retrieval, pp. 411–416. ISMIR, Kobe (2009).
11. C. Laurier, J. Grivolla, P. Herrera.: Multimodal Music Mood Classification Using Audio and Lyrics. In: 7th International Conference on Machine Learning and Applications, pp. 688–693. ICMLA, San Diego (2008).
12. Z. Sicheng, Y. Hongxun, S. Xiaoshuai, X. Pengfei, L. Xianming, J. Rongrong.: Video Indexing and Recommendation Based on Affective Analysis of Viewers. In: 19th ACM International Conference on Multimedia, pp. 1473–1476. MM, Scottsdale (2011).
13. 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), 1783–1794 (2006).
14. H. Kenmochi, H. Oshita.: VOCALOID – Commercial Singing Synthesizer Based on Sample Concatenation. In: 8th Annual Conference of The International Speech Communication Association, pp. 4009–4010. INTERSPEECH, Antwerp (2007).
15. T. Kudo, K. Yamamoto, Y. Matsumoto.: Applying Conditional Random Fields to Japanese Morphological Analysis. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 230–237. EMNLP, Barcelona (2004).