Poster: Piece – Towards Personalized Music Video Annotations Based on the User's Physiological Data

Ryan Mannschreck
Keio Media Design
Keio University
Yokohama, Japan

Kai Kunze
Keio Media Design
Keio University
Yokohama, Japan
kai@kmd.keio.ac.jp

UbiComp/ISWC'18 Adjunct, October 8–12, 2018, Singapore, Singapore
ACM 978-1-4503-5966-5/18/10.
https://doi.org/10.1145/3267305.3267656

Abstract
Our overall goal of our is to provide a personalized method to categorize and find media content of interest based for individual users, especially focusing on implicit feedback (facial expressions, posture and other reactions). This paper presents an initial study to understand tagging and annotation processes for music videos, focusing on Korean Pop videos. We present an initial experimental setup of 10 users watching 5 selected K-Pop Videos. We obtained words, key terms, and phrases users use to describe/search for the music video content in question and recording eye gaze as well as body posture and facial expressions for the participants. In addition we explore Tag identification associated with the video content in a study, to look into cultural and individual differences.

Author Keywords
Music, Posture

ACM Classification Keywords
H.5.5 [**Information and Presentation**]: Sound and Music Computing.

Introduction
There are a lot of different retrieval systems for music and other types of media content. Most of those systems work with user generated tags and detailed analysis of the con-

Figure 1: P3 displaying negative/passive posture, and P4 displaying positive/active posture

tent (e.g. parsing the songs, applying machine learning techniques to images to detect objects). We believe we can add another dimension to these systems by individualising the search queries and contents based on the users immediate physiological feedback while watching the media content or seeing search results, providing more personalized and immersive music experiences [2].

We pick Korean Pop Music videos as a starting point for our studies, as they usually provide very strong reactions to viewers not familiar with them[3]. We plan to extend the range adding music video search ability by integrating tags inspired by the visual elements of videos, user generated tags, and automatic tagging generated from body posture/facial expression.

In this paper we present initial findings form an experimental setup where we had 10 users with different cultural backgrounds watch Kpop music videos. We recorded their posture and facial expressions over a camera as well as eyegaze. We regarding cultural differences related to tagging and descriptions of KPop Videos. We conducted a preemptive study to gain information on people's uptake from music videos, what they enjoy about them, how they describe them, and how they go about finding them. From there we developed a tag gathering study, conducted in two languages to gather a baseline of tags for an annotation system. We use this information to begin developing the augmentation tool for music videos, Piece.

Approach

This work is a grounded theory study. Beginning with an initial exploratory observational research study to look into current annotation systems. Followed by an observational research study done on the common search and consumption habits of music video's consumers. Using these exploratory searches as a jumping off point, the next stage was that of a problem orientated qualitative research method using cross sectional studies consisting of descriptive surveys and interviews before and after witnessing the music video samples. We used the vehicle of Kpop music videos as a means to begin study. This way participants focused on the visual qualities rather than the lyrical ones.

The tags where gathered from users, lyrics, vocal annotation, posture, facial expression, and pre-existing systems to form a pool of data. These sources where organized into hierarchal layers to build a top-down system of categorization. Layer one saw that each music video from the sample was categorized by combining music, movie, and TV genres to create holistic genres that music videos can be placed under giving credit to its dual media style. Layer two, incorporated existing annotation systems practices. Layer three recorded what people remember, look for, and enjoy in music videos. To begin addressing cross-cultural needs we conducted layer three in two languages, English and Japanese. The answers provided by the participants in each language would then be combined and translated into each language allowing for an expanse of tagging terminology for each culture. One facet that makes this annotation system somewhat unique is its ability to take real time recordings of users while watching music videos and mark moments they respond favorably to, as shown by their postures and facial expressions. These marks are then automatically generated into tags and placed into Pieces system.

Experimental Setup

The study ran with N=10. All participants had little to no exposure to Korean pop music. None of the participants could understand Korean. The relevant information for each participant from the exploratory study is displayed in Table **??**.

Goals of the Tag Identification Study
Obtain words, key terms, and phrases users use to describe/search for music video content. Built from the interview questions asked in the first experiment, the new survey would emphasize the participant's use of adjectives, mood, and themes to describe the videos. The questions were directed by the information found in the exploratory research and from the common search engine questions users use to search for music videos online. Two, this experiment would gather data in multiple languages, English, Korean, and Japanese, as a step toward globalization of the tags. Three, responses gathered in this experiment would be added to existing data, to each language's responses, and to the vocal annotations. Addressing the duality of music videos by combining data from the three layers of data decided upon, metadata, visual descriptors, and vocal annotation.

Name	Age	Nationality	Gender	Sexuality	Languages
P1	26	Italian	Female	Heterosexual	Italian, English, Spanish
P2	25	Japanese	Male	Heterosexual	Japanese, English, German
P3	23	Indian	Male	Heterosexual	Hindu, English, Japanese
P4	26	Thai	Female	Heterosexual	Thai, Japanese, English
P5	23	Chinese	Female	Bisexual	Chinese, Japanese, English
P6	26	Thai	Female	Heterosexual	Thai, English
P7	27	Chinese	Female	Heterosexual	Chinese, Japanese, English
P8	25	Chinese	Female	Heterosexual	Chinese, Japanese, English
P9	26	Italian	Female	Heterosexual	Italian, English
P10	27	Mexican	Female	Heterosexual	Spanish, Japanese, English

Table 1: Study Participants

Participants were asked to watch 5 videos. The videos were selected from a pool of 10 videos, utilizing the Latin square method so as to ensure controlled randomization of the content. After watching each video participants were interviewed for qualitative data. Questions consisted of asking the participants for adjectives or phrases they would use to describe the music video, the movie genre they would give to the video, the music genre they would give to the video, what they believed the theme of the video was, and what they thought the key points of the video were. During the experiment, participants where videoed and the recordings saved. The videos where used to examine physiological responses by utilizing the open source program, Openpose, for analysis. To insure a baseline, participants where asked to sit for 5 minutes while recording their posture at rest. Their eye movements where also tracked using the Tobii eye tracking system to recorded visual interest.

Afterward an exit interview was conducted for more general thoughts on the music videos as a whole. A key part of the questioning was what part of the story they most remember and in which part they thought was the most important. This is significant not just because it registers their opinion but, because we can compare their responses to the physiological information we developed to see if there is a correlation of any significance. Did they show more physical reactions towards the same moments that they described as the most interesting? Did their eyes widen or did their smile increase when they viewed this particular section? The Openpose software works by utilizing the camera built into the computer. The software allows you to map key points on the body and face and follow their changes over time.[1]. By calculating the difference between each changing point for each key frame one can see how the body and face changes over time. This allows one to map, for example, the upturning of the corners of the mouth that occurs when someone smiles. By matching these body posture and facial expression changes, it is possible to see what peoples' reactions are to particular moments in the music videos physiologically.[4]

The process for this study was comprised of four parts. Step one asked participants to fill out the Positive and Negative Affect Schedule or PANAS-SF for short. This test is

Tag Identification Study Summary
We had twenty-two English-speaking and five Japanese-speaking participants. In the English-speaking group 77.3 percent where female, 18.2 percent male, and 4.5 percent answered as other. Heterosexual dominated the sexuality category with 86.2 percent of participants, 13.6 percent identified as LGBTQ. Ages ranged from 21 years old to 58 years old, with the highest result 22.7 percent being 22. Nationalities showed a wide diversity. Musical tastes spanned a wide distribution.Almost all participants find their music online, through music services or corresponding apps. Other methods for finding music mentioned where through friends/family, radio or TV, and social media. When asked if it was important to them that the language of a song was one they could understand 13.6 percent said yes, 54.5 percent do not mind at all, and 31.8 percent said sometimes it matters.63.6 percent of participants watch music videos sometimes. 22.7 percent do not watch them at all and 13.6 percent watch them frequently.

made up of two mood scales that measure positive affect and the other measuring negative affect. The scale is used to share relations between positive and negative affect and changes in mood due to external stimuli. Step two had participants fill out a survey to gather demographic info, music video consumption habits/tastes, and their exposure to Kpop. In part three participants where asked to watch five Kpop videos one at a time.After each video participants would fill out the survey for that video. They where asked to answer right afterwards so as to respond when the memory was still fresh. Questions this time focused more on the moods, aesthetic styles, and the memorable qualities of the content. The focus became more on language and the descriptors it would bring forth. At the end of the study the participants where asked to fill out the PANAS-AF scale for a second time. Reporting how they felt at the end of the experiment after being exposed to the content. The two tests where totaled up and the scores were then compared to each other. Looking to see if there was any positive or negative change to disposition after participating in this study. Users where recorded while watching the videos and run through Openpose.

Discussion and Conclusion

Initial analysis showed that the physiological information to locate key points of interest can work well. These key points of interest did show correlation to the verbal information responses given in interview questions. It was possible to see positive physical responses to portions of the content as well as negative or "bored" responses. sOften when a participant named the key points of interest or enjoyment for them it was reflected in the physiological data. The same can be said for the negative physiological responses. When participants stated distaste for a portion of a video it was clearly displayed in the physical data. The hardest state to record though was boredom. As this state of being could also be interpreted as being at rest or simply not feeling an overt feeling in general. The baseline recording was essential for understanding each person's responses. It was clear while going through the recorded point data that people express themselves in varying degrees that can only be slightly compared to the "average" human response catalog.[4] It would require more in depth processing and more precise tools of recording and measure to insure any significant data. Participants reported difficulty in categorizing the music video's genre both musically and visually.It was easier for the other videos when they where given a list of music and movie genres to select from. Even with the list though there was variation in the answers given by participants. Identifying adjectives, key terms, and the color palette for each video proved easier then picking a genre for participants. Adjectives and key terms showed a great level of over lap. Displaying a large number of similar word choices or similes.

Acknowledgments

We want to thank Prof. Juhan Nam and Chaelin Park from KAIST for their support and help. This work is also partly supported by JST CREST Grant No. JPMJCR16E1.

REFERENCES

1. Zhe Cao et al. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.
2. Haruna Fushimi and et. al. atmoSphere: designing cross-modal music experiences. In *ACM SIGGRAPH 2017 EM*.
3. Jeremiah Estela Magoncia. 2014. *OMG! Reaction Videos on YouTube: Meanings to Fandom and to K-Pop Community*. Ph.D. Dissertation.
4. Caifeng al. ets Shan. 2009. Facial expression recognition based on local binary patterns. *Image and Vision Computing* 27, 6 (2009), 803–816.