Music Video Clip Impression Emphasis Method by Font Fusion Synchronized with Music

Kosuke Nonaka(✉), Junki Saito, and Satoshi Nakamura

Meiji University, Nakano 4-21-1, Tokyo, Japan
`electro0701@gmail.com`

Abstract. Video content that displays lyrics synchronized with the music in, for example, karaoke or music videos on the Internet is becoming popular. The font can change the impression the users get from the content; therefore, by changing the font according to the content, it is possible to improve the content and the overall experience. We aim to develop a new lyric expression method by blending the existing fonts and synchronizing them with the music. In this study, we propose a method that generates fonts that match music videos by mixing fonts according to the music characteristics. Also, we reveal through experiments whether displaying lyrics in a blended font can emphasize music video impressions compared to displaying them in the existing font.

Keywords: Font · Music video · Lyric video · Emphasis impression · Subtitle · Font blending

1 Introduction

The content of music lyrics is very important. The function of lyrics is to convey information about thoughts, situations, and scenes to the listeners. Nowadays, there are many ways to read the lyrics, for example, on lyric cards in CDs, in subtitle-guided singing at karaoke clubs, or during music programs on TV.

In recent years, video with lyrics software, which uses the lyrics as the main content and synchronizes them with the music, has become popular. In Japan, there are many lyrics videos, for example, Hold Your Hand (Perfume), Aruku Around (Sakanaction), and Go Go Ghost Ship (Kenshi Yonezu). It is not unusual to see lyrics displayed in music videos created with Vocaloid, voice synthesizer software, because synthetic voices are harder to understand than human voices. Other technologies that support musical text expression are TextAlive [1] and LyricSpeaker [2]. Visualization of lyrics synchronized with the music is an effective way to understand the meaning of lyrics. This visualization provides the user with a fascinating and enjoyable experience.

The font of the lyric texts created by software such as Lyric Video or TextAlive is usually the same throughout the music video regardless of the genre, melody, and content of the lyrics. Lyrics shown as a guide in karaoke also use the same font. On the other hand, on posters and in comics, creators choose appropriate font designs to emphasize the content; they would not use a joyful font for a sad scene in a comic

© IFIP International Federation for Information Processing 2019
Published by Springer Nature Switzerland AG 2019
E. van der Spek et al. (Eds.): ICEC-JCSG 2019, LNCS 11863, pp. 146–157, 2019.
https://doi.org/10.1007/978-3-030-34644-7_12

book. Several studies have investigated how the font design can change the impression of the content. We expect that the comprehension of the music, as well as the visual and auditory experience, can be augmented if an appropriate font design is selected to display the content. As a result, kinetic text expression created with, for example, LyricSpeaker or TextAlive may become richer. Specifically, it is thought that music video entertainment will be further enhanced by making the lyrics more enjoyable.

2 Related Work

Several studies have proposed various ways to extend the viewing experience by adding other types of stimulation, such as, using vibrations synchronized with the music in an impression emphasis method [3]. Similarly, in our research, we aim to extend the visual and auditory experience by adopting an appropriate font for a music video.

A number of studies have reported that the font designs affect the impression of the content. The visual effects of the shapes of Japanese characters on printed matter have been investigated, and it has been verified that different font designs change visual impressions. Mackwiewicz et al. [4] examined the relation between the font design and human recognition. In the same study, 15 types of fonts were evaluated for 10 types of attributes, and it was clarified that the visual characteristics and design characteristics of the font were related to each other. Also, Caldwell [5] analyzed the emotional response to the Japanese font systematically and revealed a relation between the visual characteristics of the font and the emotional response.

Several studies have investigated how the impressions of the font affect human recognition or emotions. Doyle et al. [6] reported that an appropriately selected font for a product can change human behavior. He focused on the consistency of the font and product because the impressions gained from the font design affect human emotions before the meaning of the words does. Velasco et al. [7] revealed that the shape of the font and the sense of taste are related. For example, angular fonts are associated with “sourness”, “bitterness”, and “spiciness”, while rounded fonts are associated with “sweetness”. Karnel et al. [8] studied the relation between product packaging and the consumer’s behavior: a consumer with healthy habits tends to perceive a lightweight font as healthy, so this font increases the consumer’s intention to buy merchandise.

In a study of the creation of new fonts, Suveeranont et al. [9] proposed a system for creating a new font by blending a base font with any other font. However, the system requires manual corrections when the output font is distorted due to blending fonts with extremely different shapes. Contrary to this, Campbell et al. [10] proposed a method that places fonts on a two-dimensional map so that fonts with similar shapes are close to each other and no distortion occurs when they are fused. Our method also blends some existing fonts. However, unlike in Campbell’s method, we expressed the font using a numerical formula as done by Saito et al. [11] and generated blended fonts by weighting and averaging the fonts.

3 Proposed Method

In this study, we proposed a method that generates fonts that reflect music video impressions by blending existing fonts. We used six music video impressions as input: five impression classes from MIREX [12] and "cuteness" from Yamamoto et al.'s study [13]. The descriptions of C1–C6 are given in Table 1.

Table 1. Impression classes

Name	Adjectives
C1 (grandness)	Grand, sonorous, massive
C2 (vigorousness)	Lively, exciting
C3 (sadness)	Painful, sad
C4 (violence)	Aggressive, violent, exciting, passionate
C5 (funniness)	Unique, funny, roguish, whimsical, hilarious
C6 (cuteness)	Cute, lovely, adorable

We applied Saito's method of blending fonts [11]. In this method, fonts are blended by considering strokes and diameters as numerical formulas. We supposed that letters can be represented by the locus of a circle with a changing diameter and express these letters by drawing the trajectory of the circle with its center mapped on the stroke line. Specifically, as shown in Fig. 1, first, we obtained a set of circles for drawing the font, and then, to express the font smoothly, we performed cubic spline interpolation to connect the obtained set of circles as much as possible. After that we generated a circle that fills in the gaps. Next, we calculated the numerical formula, which is a sequence of the centers of the circles, by Fourier series expansion. In Fig. 1, (f(t), g(t)) is a point on a stroke of a character of a font, and the distance between the edge of the font and the point is h(t): distance information.

Each calculated stroke can be expressed as a numerical formula of t:

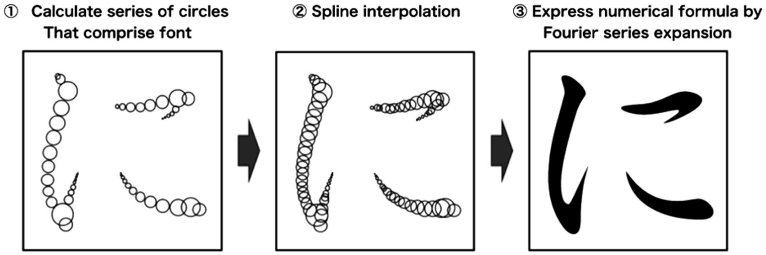
$$\begin{cases} x = f_i(t) \\ y = g_i(t) \\ r = h_i(t) \end{cases} \qquad -\pi \leq t \leq \pi$$

By fusing N fonts using the optional rate $\alpha_1 - \alpha_n$ the numerical formulas of the blended strokes are as follows:

$$\begin{cases} x = \sum_{i=1}^{N} \alpha_i f_i(t) \\ y = \sum_{i=1}^{N} \alpha_i g_i(t) \\ r = \sum_{i=1}^{N} \alpha_i h_i(t) \end{cases} \qquad \sum_{i=1}^{N} \alpha_i = 1$$

By changing the rate of blending, we can create appropriate fonts. Although the number of impression words when blending is two, we extend the number of words by fusing the aforementioned impression classes C1–C6. In the following section, we describe a preliminary survey for computing the C1–C6 impression values of fonts and music videos.

In our method, we blended fonts by associating music video impressions with font impressions. We conducted preliminary surveys on the impressions of fonts and music videos. The first survey was on the fonts. Nineteen participants evaluated their impressions of 14 fonts (Fig. 2); eight fonts were taken from previous research [11], and six fonts were selected by the authors. We showed text in each font on the web system used for the experiment. In this survey, the participants evaluated each impression class on a scale from −2 to +2 using adjectives and negative words.


① Calculate series of circles That comprise font
② Spline interpolation
③ Express numerical formula by Fourier series expansion

① $(x, y, r) = \{(x_1, y_1, r_1), (x_2, y_2, r_2), \cdots, (x_n, y_n, r_n)\}$

② $(x', y', r') = \{(x_1, y_1, r_1), (x_{1,1}, y_{1,1}, r_{1,1}), \cdots (x_{1,m}, y_{1,m}, r_{1,m}), (x_2, y_2, r_2), \cdots, (x_n, y_n, r_n)\}$

③ $X_1 = f_1(t), Y_1 = g_1(t), R_1 = h_1(t)$
$X_2 = f_2(t), Y_2 = g_2(t), R_2 = h_2(t)$
$X_3 = f_3(t), Y_3 = g_3(t), R_3 = h_3(t)$

Fig. 1. Steps of formulating fonts

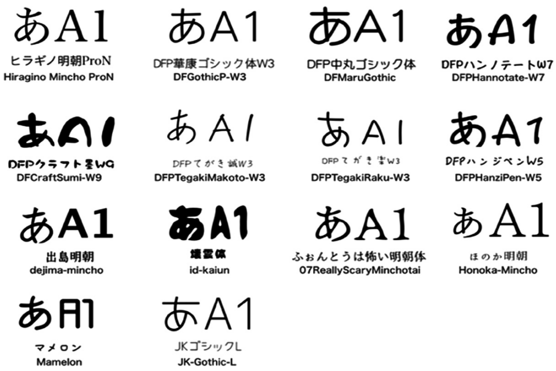

ヒラギノ明朝ProN
Hiragino Mincho ProN
DFP華康ゴシック体W3
DFGothicP-W3
DFP中丸ゴシック体
DFMaruGothic
DFPハンノテートW7
DFPHannotate-W7
DFPクラフト墨W9
DFCraftSumi-W9
DFPてがき誠W3
DFPTegakiMakoto-W3
DFPてがき楽W3
DFPTegakiRaku-W3
DFPハンジペンW5
DFPHanziPen-W5
出島明朝
dejima-mincho
壊雲体
id-kaiun
ふぉんとうは怖い明朝体
07ReallyScaryMinchotai
ほのか明朝
Honoka-Mincho
マメロン
Mamelon
JKゴシックL
JK-Gothic-L

Fig. 2. Used fonts

Next, we conducted a survey on the impressions of the music videos. Nineteen participants (university students aged 19–23) watched 18 shortened music videos and evaluated six classes of impressions. The duration of each shortened music video was about 20–30 s. In this evaluation, each impression class was evaluated on a scale from −2 to +2 using adjectives and negative words. Also, the order of the music videos was made random for each participant.

4 Experiment

Experiment procedure: We hypothesized that the impression of the music video can be emphasized by using a blended font for lyrics. We conducted experiments to compare a blended font and two fonts created under two conditions.

In our experiment, we verified the usability of a font created by our method by watching a music video that was synchronized with lyrics created in three ways. For generating fonts, we also used two other methods in addition to our method.

1. Blended font: A font generated by adopting the top four neighboring fonts obtained from the impressions of a target music video, weighted averaging, and fusing.
2. Neighborhood font: A font determined to be the closest to the impression of a target music video.
3. Impression_0 font: A font generated by the font fusion method with all six impression class inputs as 0.

In the experiment, the choruses of the same 18 music videos from the preliminary survey were used. A total of 54 types of music videos were generated, and three types of fonts were superimposed on the lower part of the screen. Although there may be important factors for displaying lyrics such as color, size or position, we unified these conditions because subtitles are static throughout the music video in many cases. The generation system was implemented using Processing, open-source graphical library. An example of the generated font is shown in Fig. 3. And examples of the fonts and each impression class are shown in Fig. 4.

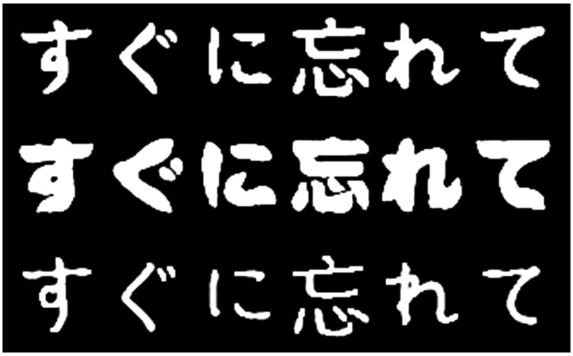

すぐに忘れて
すぐに忘れて
すぐに忘れて

Fig. 3. Fusion example of each font. From top: fusion font, neighborhood font, and impression_0 font.

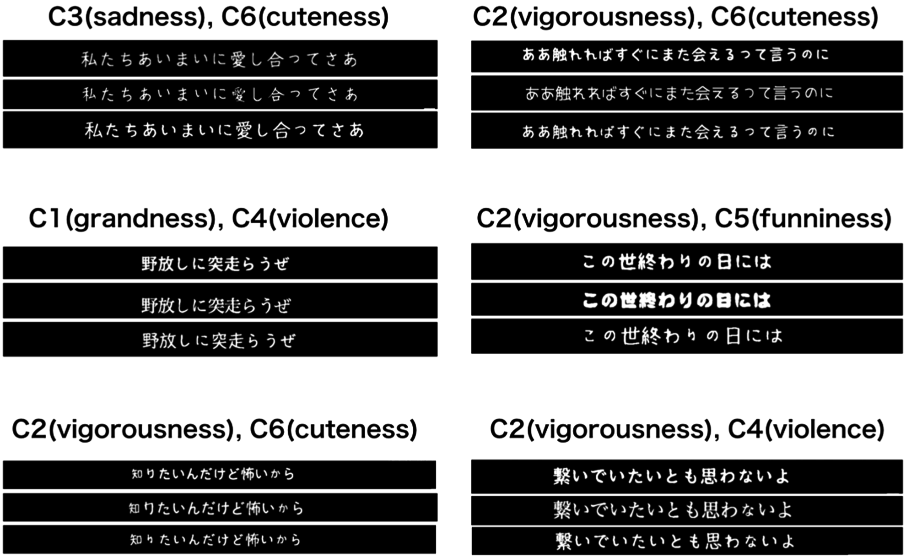

C3(sadness), C6(cuteness)
私たちあいまいに愛し合ってさあ
私たちあいまいに愛し合ってさあ
私たちあいまいに愛し合ってさあ
C2(vigorousness), C6(cuteness)
ああ触れればすぐにまた会えるって言うのに
ああ触れればすぐにまた会えるって言うのに
ああ触れればすぐにまた会えるって言うのに
C1(grandness), C4(violence)
野放しに突走らうぜ
野放しに突走らうぜ
野放しに突走らうぜ
C2(vigorousness), C5(funniness)
この世終わりの日には
この世終わりの日には
この世終わりの日には
C2(vigorousness), C6(cuteness)
知りたいんだけど怖いから
知りたいんだけど怖いから
知りたいんだけど怖いから
C2(vigorousness), C4(violence)
繋いでいたいとも思わないよ
繋いでいたいとも思わないよ
繋いでいたいとも思わないよ

Fig. 4. Generated font of lyrics and their impression classes. In each music video, from top: fusion font, neighborhood font, and impression_0 font.

Nineteen participants (university students aged 19–23) watched music videos with superimposed lyrics in fonts created under three conditions. Although some of the participants had participated in the data set construction, it was judged to have no influence on the experiment because the font design, music animation and the superposition were considered to be independent from each other. For the experiment, we prepared a web system, which was used to evaluate six impression classes after viewing music videos on a five-point scale steps from −2 to +2 (Table 1).

Also, even if the impression is emphasized, it is conceivable that the character design would not match the mood of the music video. Therefore, to quantify the degree of matching between the font and the music video, we asked the question “Did the font match the music video?” The answer scale had five steps from −2 (“It did not match at all”) to +2 (“It matched very well”). In addition, the order of music videos was made random for each participant; therefore we concluded that there was no order effect.

4.1 Results

Figure 5 shows the degree of matching when the fonts that were created using three different methods were applied to the lyrics of the music videos. Seven music videos were evaluated to be most suitable for the fusion font, nine music videos were evaluated as most suitable for the neighborhood font, and two music videos were evaluated as most suitable for the impression_0 font. From these results, we can assume that a font synchronized with an impression is suitable regardless of fusion, although the most suitable font to be used for the lyrics depends on the music video. Music videos C, F, I, and R had an overall low degree of matching.

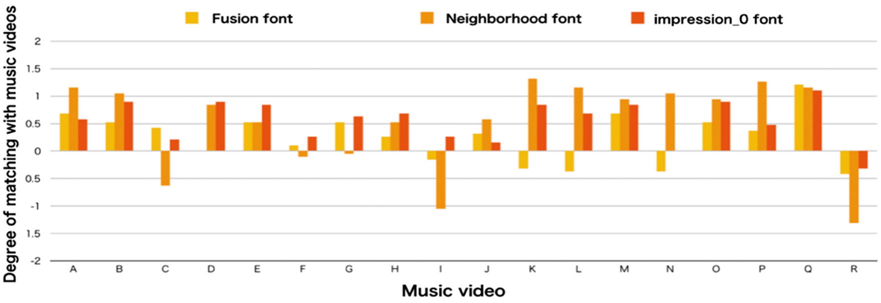


Fig. 5. Matching degree of music videos and lyrics for each font.

Figure 6 shows a list of comparisons of the values from the preliminary survey and the experiments for each music video. The blue columns show the values of the preliminary survey, and the green, yellow, and red columns are the values obtained in the experiments. It can be seen that there is variation in impression emphases. The number of evaluation values of any one of the three types of fonts exceeding the value of the pre-survey is 8 for C1, 8 for C2, 14 for C3, 6 for C4, 11 for C5, and 15 for C6. This shows that the impressions of C3 (sadness), C5 (funniness), and C6 (cuteness) are well emphasized.

4.2 Analysis

A numerical comparison is performed for each impression class of each music video. From Fig. 6, it can be seen that C3 and C4 in the impression class have a large variation in values for each font, and impression emphasis did not occur as intended. Figure 7 shows graphs in which the impression value data of the font created in the preliminary survey are sorted in ascending order for each of the C1 to C6 impression classes. This graph indicates that C3 and C4 are less than 0, which is the median value in the questionnaire survey, for most of the fonts. In other words, among the impressions of music videos, those with higher C3 and C4 are considered to have the same font each time and do not reflect the impression of the music videos accurately. In addition, the variation in the values of each font is small, and it is considered that a similar font has been generated as a result of font fusion. Furthermore, although the lyrics were added to the lower part of the music video screen this time, if the participants were familiar with the music, it is possible that they could have watched the music video without looking at the lyrics. In the future, we plan to conduct an experiment that takes into consideration the way the viewer sees the subtitles, for example considering karaoke lyrics and TV subtitles.

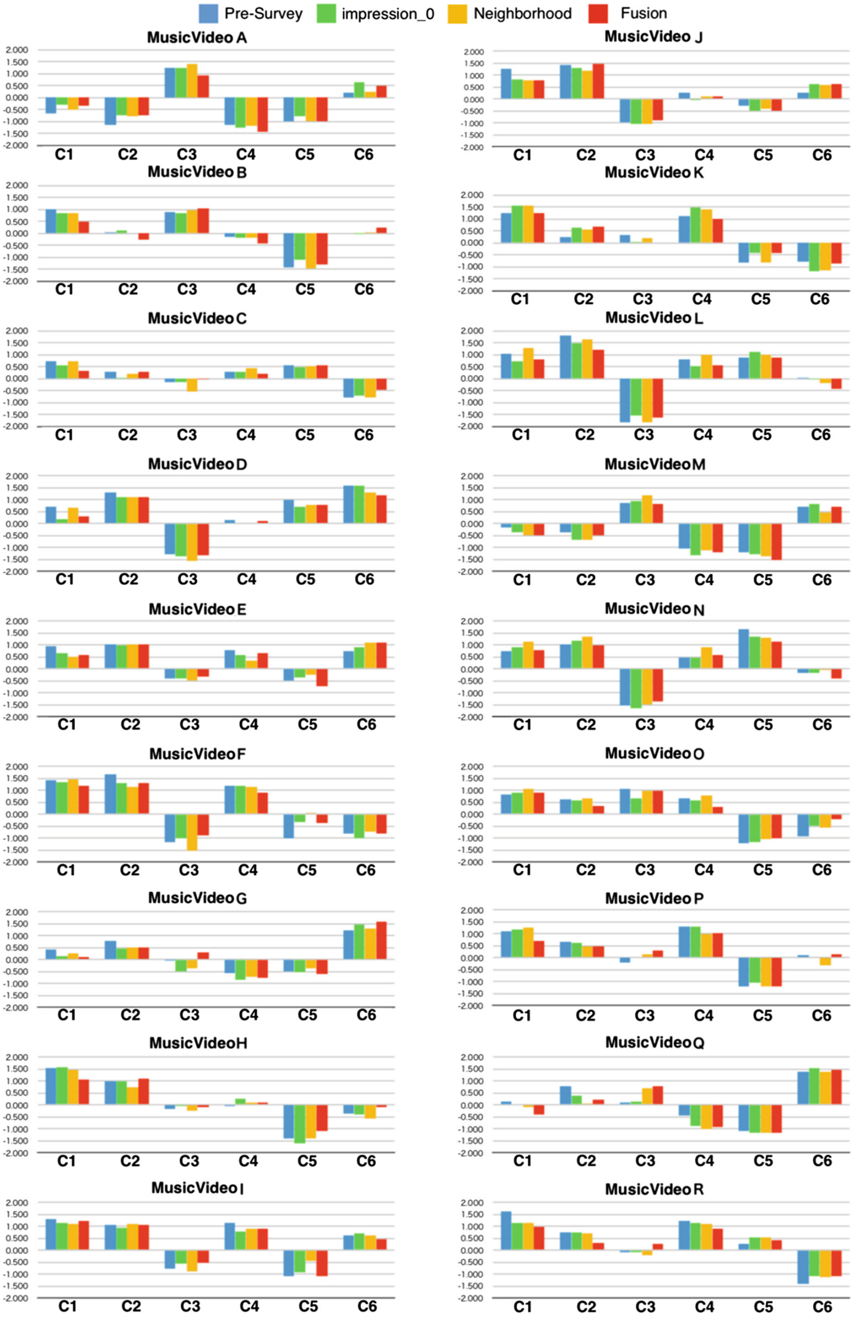


Fig. 6. Pre-survey values and experimental results for each music video.

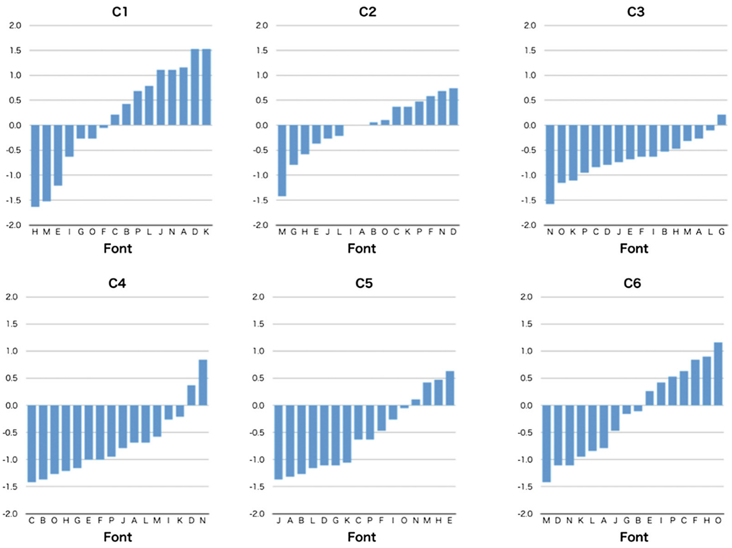


Fig. 7. Matching degree between music videos and lyrics for each font

Next, we focused on the differences in impression emphasis among fonts. Figure 8 shows that for each music video, there are differences between the fusion font and the impression_0 font and between the neighboring font and the impression_0 font. The numbers are graphed for each numerical interval. From Fig. 8, it can be seen that in music videos F, H, I, L, N, O, and R, the neighborhood fonts are evaluated higher than fonts created by the proposed method. By looking at music videos F, H, I, L, N, O, and R in Fig. 6, it can be seen that all values of C1, C2, and C4 in the preliminary survey are high. C3, C5, and C6 represent "sadness", "funniness", and "cuteness" and can convey the impression from the font. However, C1, C2, and C4, which represent "grandness", "vigorousness", and "violence", represent emotions that may be difficult to convey by static characters. In other words, when expressing C1, C2, and C4 as characters, it is better to animate the font rather than focus on the font design.

In our method, four fonts are fused, and it is possible that an inappropriate font may be generated if the degree of matching with the music video is not very high. Therefore, to analyze with high degree of matching, we first selected fused fonts and neighborhood fonts that had a high degree of matching and a difference compared with the impression_0 font. Then, we calculated the difference in size of the impression vector between the fusion font and impression_0 font and between the neighborhood font and impression_0 font. Figure 9 shows these values as a graph. When the degree of matching between the presented font and the music video is +2 (very well matched), the fusion font can better emphasize the impression compared to the neighborhood font. However, it can be confirmed that the fusion font is not effective when the degree of matching is +1 (matched). Therefore, it is possible that the font fusion method can greatly emphasize the impression if the degree of matching with the mood in the music video is high. Otherwise, the impression will be suppressed.

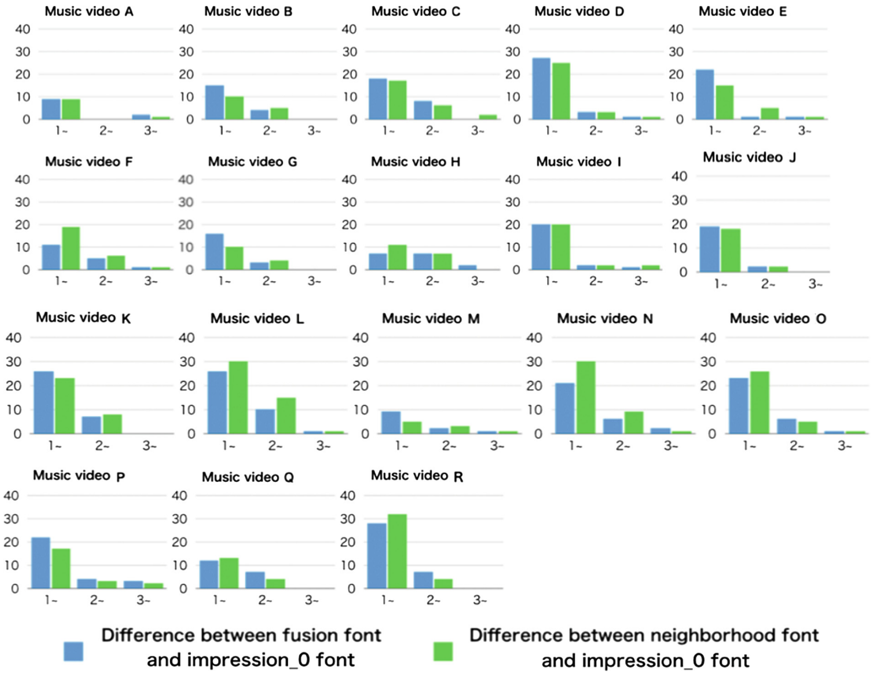


Fig. 8. Distribution of differences for each music video.

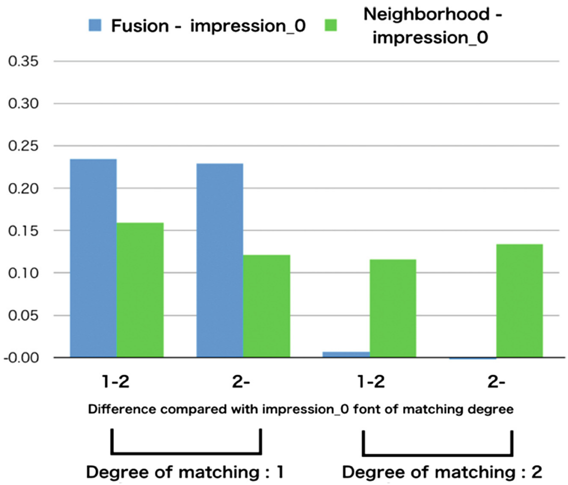


Fig. 9. Difference in size of impression vectors.

5 Conclusion and Future Work

In this research, we proposed a method to dynamically generate fonts suitable for music videos by fusing existing fonts for music videos. A preliminary survey on fonts and music videos was conducted, and later fonts were fused based on the data received from the survey. When a user views a music video, we compared impression values when applying the font created by the proposed method, font closest to the impression, and font with the impression value set to 0 to the lyrics.

We conducted experiments to verify whether the font emphasizes the impression of the music videos. As a result, we confirmed that the proposed method is useful for music videos with high impression values of C3 (sadness), C5 (funniness), and C6 (cuteness). On the other hand, in the case of music videos with high impression values of C1 (grandness), C2 (vigorousness), and C4 (violence), it is difficult to emphasize the impression with static characters. In addition, it seems that the degree of matching with the music video is greatly affected by the impression emphasis when using the fusion font.

In the future, this method will make users feel more empathy with the music video while watching it. For example, we believe that the way people sing at karaoke would reflect their impression of the music video of the song. In addition, it is possible to fill the gap between the impression of the music that the composers tried to express when they composed it and the impression that the viewers receive. As described above, we consider this method is not only to convey the content of the lyrics, but also to make the lyrics more enjoyable and to contribute greatly to the creation of highly entertaining music and video content.

The limitations of this research are that the impression values of the selected font were uneven, and there was a possibility that the participants viewed the music videos without looking at the lyrics. Therefore, in the future, we plan to have a pre-selection stage to consider the variations in font impression values and then conduct experiments in which the participants watch music videos while looking at the lyrics. Also, in the proposed method, the impression emphases of music classes C1, C2, and C4 did not show the intended results, so we plan to develop an impression emphasis method that uses animation and color.

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

References

1. Kato, J., Nakano, T., Goto, M.: TextAlive: integrated design environment for kinetic typography. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 3403–3412 (2015)
2. Lyric Speaker. https://lyric-speaker.com/index.html. Accessed 16 Aug 2019
3. Yoshida, R., Ideguchi, T., Ooshima, K.: An examination of a music appreciation method incorporating tactile sensations from artificial vibrations. In: Fourth International Conference on Innovative Computing, Information and Control (ICICIC), pp. 417–420 (2009)

4. Mackiewicz, J., Moeller, R.: Why people perceive typefaces to have different personalities. Proc. IPCC **2004**, 304–313 (2004)
5. Caldwell, J.: Japanese typeface personalities: are typeface personalities consistent across culture? In: IEEE International Professional Communication 2013 Conference, pp. 1–8 (2013)
6. Doyle, J.R., Bottomley, P.A.: Mixed messages in brand names: separating the impact of letter shape from sound symbolism. Psychol. Market. **28**, 749–762 (2011)
7. Velasco, C., Woods, A.T., Hyndman, S., Spence, C.: The taste of typeface. i-Perception **6**(4), 1–10 (2015)
8. Karnal, N., Machiels, C.J.A., Orth, U.R., Mai, R.: Healthy by design, but only when in focus: communicating non-verbal health cues through symbolic meaning in packaging. Food Qual. Prefer. **52**, 106–119 (2016)
9. Suveeranont, R., Igarashi, T.: Example-based automatic font generation. In: Taylor, R., Boulanger, P., Krüger, A., Olivier, P. (eds.) SG 2010. LNCS, vol. 6133, pp. 127–138. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13544-6_12
10. Campbell, N. D.F., Kautz, J.: Learning a manifold of fonts. In: ACM Transactions on Graphics (SIGGRAPH), vol. 33, no. 4 (2014)
11. Saito, J., Nakamura, S.: Fontender: interactive Japanese text design with dynamic font fusion method for comics. In: Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, W.-H., Vrochidis, S. (eds.) MMM 2019. LNCS, vol. 11296, pp. 554–559. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05716-9_45
12. Hu, X., Downie, J., Laurier, C., Bay, M., Ehmann, A.: The 2007 MIREX audio mood classification task: lessons learned. In: Proceedings of 9th International Conference on Music Information Retrieval, pp. 462–467 (2008)
13. Yamamoto, T., Nakamura, S.: Leveraging viewer comments for mood classification of music video clips. In: Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013), pp. 797–800 (2013)