

## Article

# Style Transformation Method of Stage Background Images by Emotion Words of Lyrics

Hyewon Yoon, Shuyu Li and Yunsick Sung \* 

Department of Multimedia Engineering, Dongguk University-Seoul, Seoul 04620, Korea; hyewon@dongguk.edu (H.Y.); lishuyu@dongguk.edu (S.L.)

\* Correspondence: sung@dongguk.edu

**Abstract:** Recently, with the development of computer technology, deep learning has expanded to the field of art, which requires creativity, which is a unique ability of humans, and an understanding of the human emotions expressed in art to process them as data. The field of art is integrating with various industrial fields, among which artificial intelligence (AI) is being used in stage art, to create visual images. As it is difficult for a computer to process emotions expressed in songs as data, existing stage background images for song performances are human designed. Recently, research has been conducted to enable AI to design stage background images on behalf of humans. However, there is no research on reflecting emotions contained in song lyrics to stage background images. This paper proposes a style transformation method to reflect emotions in stage background images. First, multiple verses and choruses are derived from song lyrics, one at a time, and emotion words included in each verse and chorus are extracted. Second, the probability distribution of the emotion words is calculated for each verse and chorus, and the image with the most similar probability distribution from an image dataset with emotion word tags in advance is selected for each verse and chorus. Finally, for each verse and chorus, the stage background images with the transferred style are outputted. Through an experiment, the similarity between the stage background and the image transferred to the style of the image with similar emotion words probability distribution was 38%, and the similarity between the stage background image and the image transferred to the style of the image with completely different emotion word probability distribution was 8%. The proposed method reduced the total variation loss of change from 1.0777 to 0.1597. The total variation loss is the sum of content loss and style loss based on weights. This shows that the style transferred image is close to edge information about the content of the input image, and the style is close to the target style image.



**Citation:** Yoon, H.; Li, S.; Sung, Y. Style Transformation Method of Stage Background Images by Emotion Words of Lyrics. *Mathematics* **2021**, *9*, 1831. <https://doi.org/10.3390/math9151831>

Academic Editors:  
Ezequiel López-Rubio,  
Esteban Palomo and  
Enrique Domínguez

Received: 20 May 2021  
Accepted: 30 July 2021  
Published: 3 August 2021

**Keywords:** image style transformation; lyrics to image style; emotion; deep learning; style transfer

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Advances in computer technology have led to technological innovations such as information revolution, big data processing, and active use of networks. These innovations have increased interest in AI [1,2]. In recent years, AI has been researched in the field of art, which requires creativity, an inherent ability of humans. The field of art has been integrated with various industrial fields, such as AI, which is used in stage art in combination with stage effects. When a singer dances and sings, the audience views the singer's stage performance in combination with stage effects. Stage effects determine the stage mood using several important elements, such as lighting, music, acting, and stage background, which visually convey emotions associated with the songs to the audience. Among stage effects, stage background has been transitioning from the expression method of using props to the expression method of media performance that uses images through large light-emitting diode (LED) screens or projectors [3]. In general, background stage images used in media performances are selected in advance by professional stage designers at

the planning stage. Stage background images of the singing performance represent the emotions expressed in song lyrics. It is difficult for computers to represent emotions that humans have been manually designing for stage background images. Recently, research was conducted to enable AI to design stage background images in place of humans. In this approach, a stage background image recommendation system is used to automatically compose stage background images according to dance styles without professional stage designers. However, the limitation of the stage background images selected through conventional recommendation systems is that the emotions to be represented in the song lyrics are not reflected in the stage performance. It would be ideal to represent emotions represented in song lyrics through stage background images during stage performances. Research regarding the reflection of emotions contained in song lyrics in a stage background are scarce; however, it is possible to use research that transforms background images according to their meanings or purpose by synthesizing background images with text or images containing the meaning to be represented. There is research that partially transforms images using the content contained in text [4–6] and transfers the style such as color, line, and texture of image to another image [7–11]. The existing stage background image recommendation system recommends images for dancers, but this does not include the characteristics of the song lyrics.

This paper proposes a method to transform the multiple styles of stage background images based on the emotion words contained in each verse and chorus of song lyrics. First, the lyrics selected by the user are divided into sentences. Multiple verses and choruses are derived from the lyrics, one at a time and compared to the emotion word dictionary to extract emotion words included in each verse and chorus. Second, the probability distributions of the emotion words are calculated for each verse and chorus and the image with the most similar probability distribution from the image dataset with the emotion word tags in advance is selected for each verse and chorus. Finally, for each verse and chorus, the stage background images with the transferred style are outputted for each verse and chorus. The advantages of the proposed method are as follows.

- It uses emotion words contained in song lyrics to transform the style of stage background images. Audience immersion can be increased by using stage background images to represent emotions expressed in the song lyrics used for singing in stage performances. Emotions that are complex to represent using computers can be represented.
- Certain emotions that are difficult for humans to determine intuitively can be represented because the proposed method can transform the style of images based on an image with a high correlation with the emotion represented using lyrics.

The remainder of this paper is organized as follows. Section 2 introduces the stage background recommendation system, methods of extracting the visual features of emotion words, and image style transformation methods reported in related work, and examines their limitations and technical constraints. Section 3 describes the method proposed to derive emotion words contained in the verses and choruses of song lyrics to reflect the emotions in stage background images and apply the style that is directly related to the derived emotion words to the stage background images. Section 4 verifies whether the stage background images are transformed according to the probability distribution of emotion words represented for each verse and chorus. Section 5 summarizes the findings and describes the limitations of this research and future research directions.

## 2. Related Work

In this section, we introduce the existing stage background recommendation system, methods of extracting the visual features of emotion words and image style transformation methods. Song lyric-based style transformation methods are compared, and their limitations and technical constraints are examined.

### 2.1. Stage Background Image Recommendation System

A stage background image recommendation system recommends stage background images by reflecting the dancer's preferences and dance styles such as ballet, belly dance, street dance, modern dance, tango, and waltz [3]. Dancers choose familiar or favorite stage background images. Therefore, the stage background images can be artistic images or actual photographs that the dancers prefer. Reference [3] proposed a model that predicts users' preferred images through social media. The proposed model predicts the K number of images that the dancer (user) is most likely to use as stage background images via three procedures. First, the features of the images shared by the dancer on social media (Pinterest) are extracted. Second, the profile of the dancer is learned based on the features of the shared images. Third, the interest level of the dancer in each candidate image is predicted, and the candidate images are ranked according to the dancer's predicted interest level. However, because only dances are reflected, the stage background images from the stage background recommendation system cannot represent the emotions that a stage performance aims to express through lyrics.

### 2.2. Emotion Classification

To express emotions using images that are difficult to process using computers, research was conducted recently to improve the accuracy of the sentimental understanding of human emotions [12–18].

Human emotions are visualized and used in psychotherapy, image search, etc. In general, models for representing emotions are divided into two types: categorical emotion states (CES) models, which classify emotions into several basic categories such as fear, amusement, and sadness, and dimensional emotion space (DES) models, which use three-dimensional emotion space such as arousal, time, and harmony. As it is difficult to construct a multidimensional emotion space using information about time included in song lyrics, we used the CES method to consider images as a basic category of emotions. The CES method is easy for users to understand and convenient for emotion classification of images. The research in [12] used the CES method to extract principles-of-art-based emotion features (PAEF) to classify features of emotions included in images to understand the relationship between artistic principles and emotions. PAEF are a combination of representation features derived from the principles of balance, emphasis, harmony, variety, gradation, and movement. PAEF are used to classify the basic emotion words evoked in humans through images. A psychological research classified common basic emotion words into eight categories based on images through facial electromyography, heart rate, finger temperature, etc. That is, emotions contained in images are classified into eight categories, which define anger, disgust, fear and sadness as negative emotions and amusement, awe, contentment, and excitement as positive emotions [13]. These are called images of emotional levels, whereby an image of emotional level refers to the relationship between the style, such as color, saturation, brightness, and contrast, and the emotional effect derived from art theory [17]. The level of basic emotion words defined in the eight categories is classified for images. To evaluate this, the participants looked at the images, selected the most appropriate basic emotion category, and evaluated the emotional labels of the images. However, because it is not possible to visualize the features of images classified with the level of basic emotion words, there is no way of knowing the images that are appropriate for the stage background. Therefore, it is necessary to derive a method of visualizing the features of each basic emotion to find its relationship with the song lyrics and reflect them in the stage background images.

### 2.3. Style Transfer

Style is transferred to reflect the features of each basic emotion word in the stage background images. Usually, style transfer is used to transfer the image style. Style transfer consists of content image and style image. Content image refers to an image that has information such as an object or a common landscape that people can usually recognize,

and style image refers to an image that has information such as color or texture that will be combined with the content image. Style transfer transfers the style based on a convolutional neural network (CNN) [10,11] and a generative adversarial network (GAN) [5–7]. Style transfer based on the CNN model extracts features by separating content and style in an image. Training is performed to extract content features from deep layers and extract style features from middle layers through the CNN model. The GAN model is used to change the content in detail, but in this research, the CNN model is used because it changes the overall image style.

#### 2.4. Comparison of Methods for Image Style Transformation

Table 1 presents the difference between the existing methods of transforming image style and the proposed method. The research in Zhao et al. [12] investigates the concept of the principle of art and its effect on emotion and classifies emotion images into eight basic emotion words. However, because many images are classified for each basic emotion word, it is difficult to find an appropriate image for the stage background image.

**Table 1.** Comparison of the proposed method and image synthesis methods.

	Zhao et al. [12]	Machajdik et al. [13]	Zhao et al. [17]	The Proposed Method
Training data	IAPS, Art photo, Abstract painting	IAPS, Art photo, abstract painting	User's metadata, IAPS, Abstract painting, Flickr	Song lyrics, Flickr
Encoding	Image-based	Image-based	User's metadata-based	One-hot encoding, Texture-based
Model	SVM	Waterfall segmentation algorithm	Multi-Task Hypergraph learning	CNN

### 3. Method of Transferring Image Style Based on Song Lyrics

This section presents the proposed method to transfer stage background images using the emotion words contained in song lyrics. The proposed method consists of the lyrics preprocessing stage, which extracts the probability distribution of emotion words for each verse and chorus from selected lyrics and the emotion image processing stage, which transfers the styles of each verse and chorus images related to the extracted probability distribution of emotion words. The proposed method transfers the stage background image using styles of images related to emotion words extracted from each verse and chorus. The number of images with representative emotion image styles applied is equal to the sum of the number of verses and choruses from selected lyrics.

#### 3.1. Overview

Figure 1 is the overview of the proposed method. The proposed method is composed of the lyrics preprocessing stage and the emotion image processing stage. Table 2 is the description of all stages. In the lyrics preprocessing stage, the selected lyrics by a user are extracted into verses and choruses, and the probability distribution of emotion words contained in each verse and chorus is extracted separately. In the emotion image processing stage, the emotion images with tags, where the tags are matched to the corresponding emotion images in advance, are selected from the extracted emotion words of each verse and chorus and the stage background image is transferred to the different styles of the selected emotion image according to each verse and chorus.

**Table 2.** Process for transforming stage background image style based on song lyrics.

Stage	Description
Lyric preprocessing	User's selected lyrics are divided into verses and choruses, and a probability distribution of emotion words is extracted for each verse and chorus.
Emotion image processing	From emotion images with tags, the appropriate images are selected for each verse and chorus, and styles of selected images transferred to stage background image.

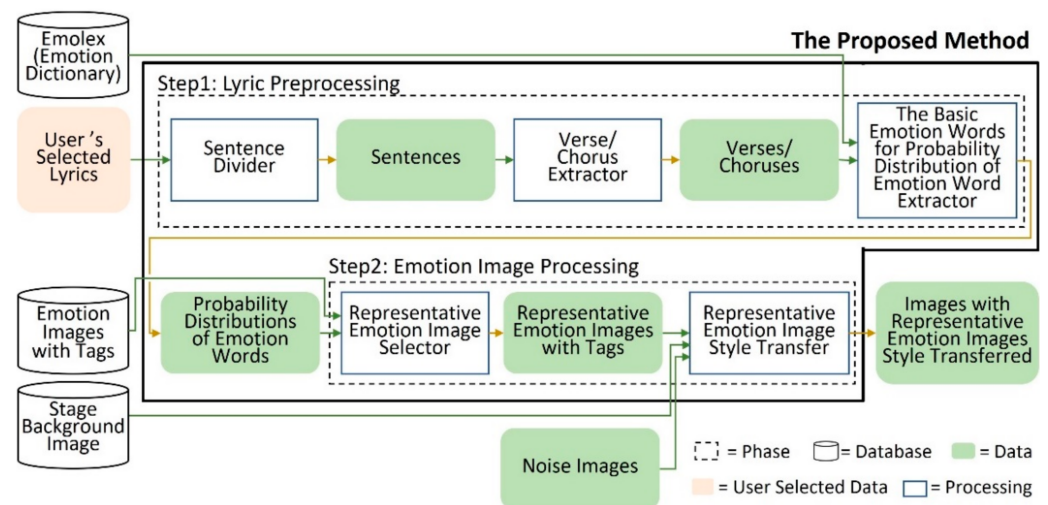


Figure 1. Process of applying emotions of song lyrics to stage background images.

### 3.2. Step 1: Lyric Preprocessing

The lyric preprocessing step is composed of a sentence divider, verse/chorus extractor and the basic emotion words for the emotion word extractor. The sentence divider divides the lyrics into sentences. The verse/chorus extractor extracts the selected lyrics into verses and choruses. The basic emotion words for the emotion word extractor extracts the probability distribution of emotion words contained in each verse and chorus. Figure 2 is an overview of step 1.

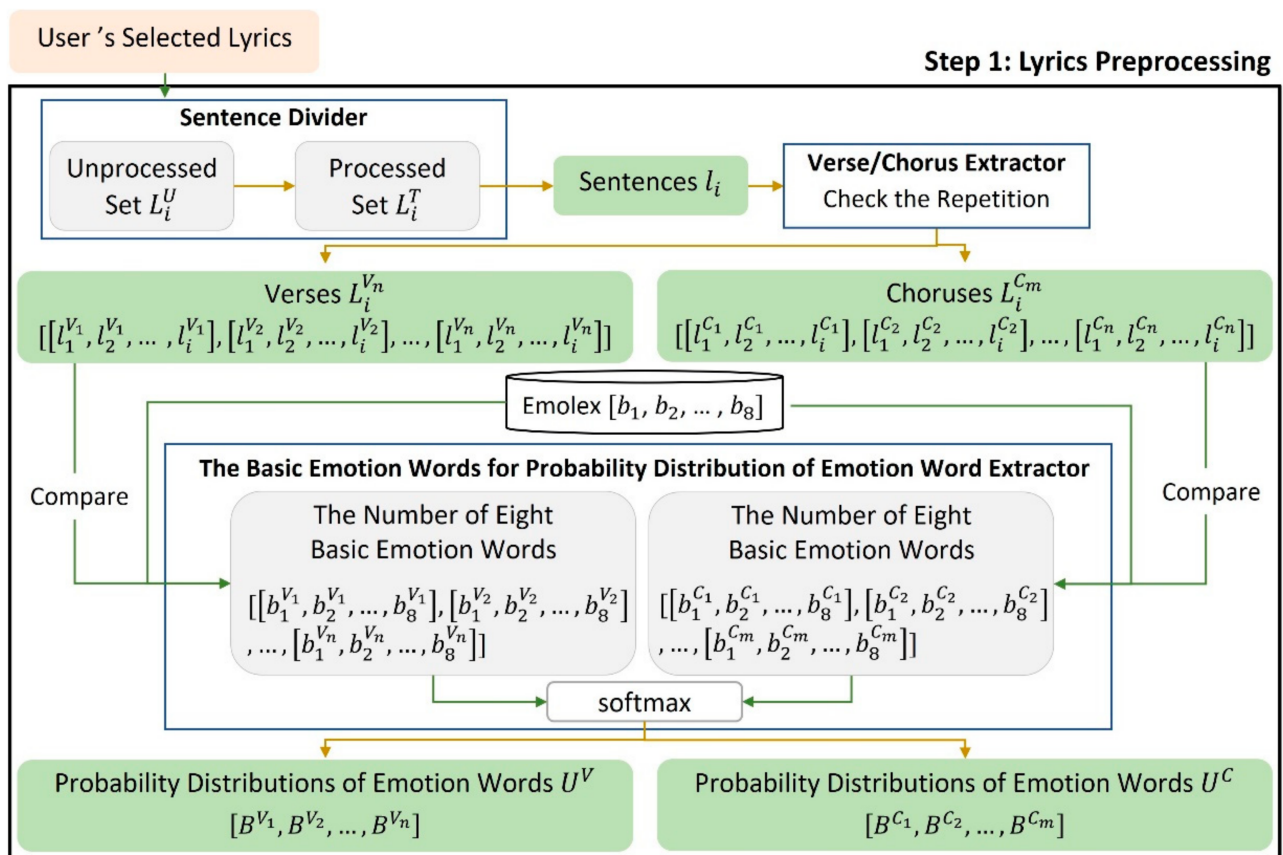


Figure 2. Step 1: Lyric preprocessing.

The sentence divider divides the lyrics into a set of sentences considering capital letters. The set of sentences in the lyrics is defined as the unprocessed set  $L_i^U$ . All sentences in  $L_i^U$  are processed as a set  $L_i^T$ , which is classified as verses and choruses through the classification process. This is repeated until there are no sentences left in  $L_i^U$  and all sentences in  $L_i^T$  are processed.

The verse/chorus extractor executes the following processes. The user's selected lyrics consist of  $n$  verses and  $m$  choruses. The  $i$ th sentence inputted in  $L_i^T$  is compared to the sentences in  $L_i^U$ , and the frequency is repeatedly checked. The set of sentences with no repetition in the lyrics as verse  $L^{V_n} = \left[ [l_1^{V_1}, l_2^{V_1}, \dots, l_i^{V_1}], [l_1^{V_2}, l_2^{V_2}, \dots, l_i^{V_2}], \dots, [l_1^{V_n}, l_2^{V_n}, \dots, l_i^{V_n}] \right]$  and the set of sentences that are repeated in the lyrics are classified as chorus  $L^{C_n} = \left[ [l_1^{C_1}, l_2^{C_1}, \dots, l_i^{C_1}], [l_1^{C_2}, l_2^{C_2}, \dots, l_i^{C_2}], \dots, [l_1^{C_m}, l_2^{C_m}, \dots, l_i^{C_m}] \right]$ .

The basic emotion words for probability distribution of the emotion word extractor compares the  $L^{V_n}, L^{C_m}$  with an Emolex (Emotion Dictionary) [19] and finds the matching emotion words. The Emolex consists of a total of 14,182 words classified into the basic emotion words, as shown in Figure 3, and information on whether they are positive emotion or negative emotion is also included. All basic emotion words are expressed by eight emotions, categorized into the positive emotions of anticipation, joy, surprise and trust, and the negative emotions of anger, disgust, fear and sadness, as proposed by Plutchik [20] to provide a high-dimensional emotion lexicon [19]. The extracted emotion words in each verse and chorus are replaced with the classified basic emotion words. Each basic emotion word is counted by the corresponding numbers, the number of anticipation  $b_1$ , that of joy  $b_2$ , that of trust  $b_3$ , that of surprise  $b_4$ , that of anger  $b_5$ , that of fear  $b_6$ , that of sadness  $b_7$ , and that of disgust  $b_8$ . The probability distribution of the basic emotion words is calculated. The set that counts the number of eight basic emotion words contained in  $L_i^T$  is defined as  $B$ . The number of eight basic emotion words included in the  $n$ th verse from  $L^{V_n}$  is stored in  $B^{V_n} = \left[ [b_1^{V_1}, b_2^{V_1}, \dots, b_8^{V_1}], [b_1^{V_2}, b_2^{V_2}, \dots, b_8^{V_2}], \dots, [b_1^{V_n}, b_2^{V_n}, \dots, b_8^{V_n}] \right]$  and the number of eight basic emotion words included in the  $m$ th chorus from  $L^{C_m}$  is stored in  $B^{C_m} = \left[ [b_1^{C_1}, b_2^{C_1}, \dots, b_8^{C_1}], [b_1^{C_2}, b_2^{C_2}, \dots, b_8^{C_2}], \dots, [b_1^{C_m}, b_2^{C_m}, \dots, b_8^{C_m}] \right]$ . The probability distributions of the basic emotion words included in verses and choruses are defined as  $U^V = \text{softmax}(B^{V_n})$ ,  $U^C = \text{softmax}(B^{C_m})$ , and calculated as Equation (1).

$$U^V = \text{Softmax}(B^{V_n}) = \frac{e^{b_i^{V_n}}}{\sum_{j=1}^8 e^{b_j^{V_n}}}, \quad i \in (1, 2, \dots, 8) \quad U^C = \text{Softmax}(B^{C_m}) = \frac{e^{b_i^{C_m}}}{\sum_{j=1}^8 e^{b_j^{C_m}}}, \quad i \in (1, 2, \dots, 8) \quad (1)$$

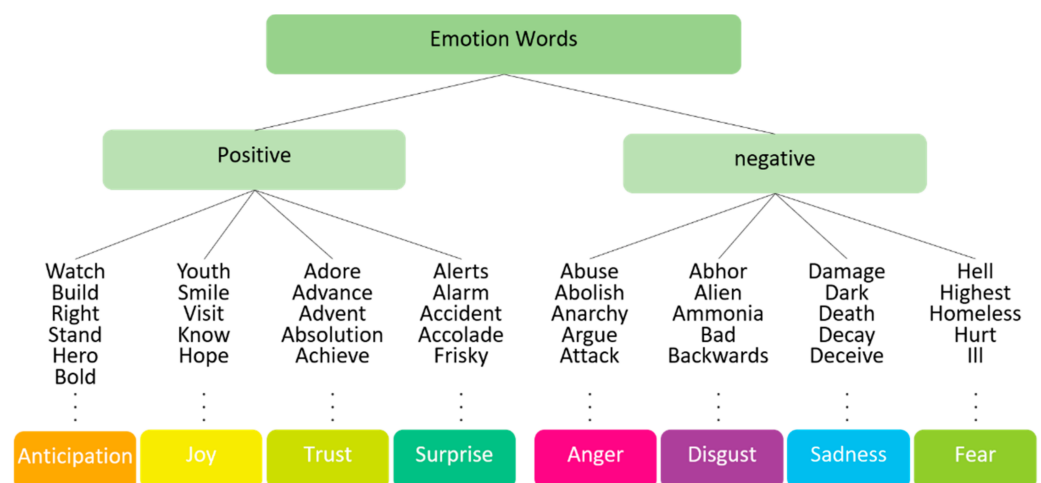


Figure 3. Classification of the emotion lexicon.

### 3.3. Step 2: Emotion Image Processing

The emotion image processing step is composed of a representative emotion images selector and representative emotion images style transfer. The representative emotion images selector searches and selects images with a probability distribution similar to the  $U^V$ ,  $U^C$  of the emotion images with tags. The representative emotion images style transfer transfers stage background images into the styles of the selected emotion images with tags. Figure 4 shows an overview of step 2.

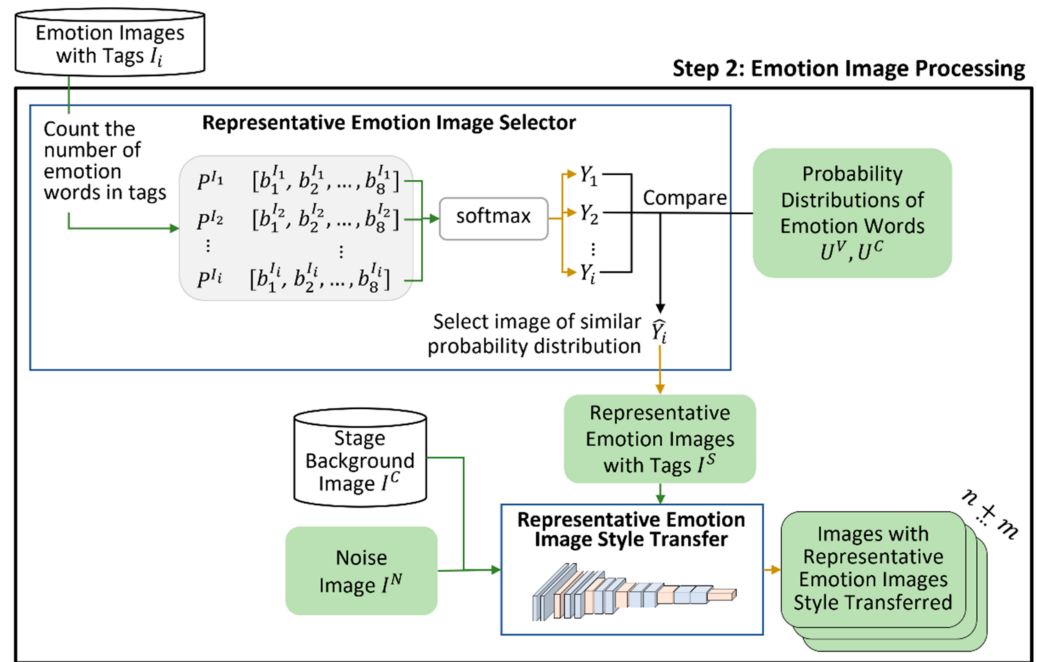


Figure 4. Step 2: Emotion image processing.

The representative emotion images selector selects emotion images with tags similar to probability distributions  $U^V$ ,  $U^C$  contained in each verse and chorus in the lyric preprocessing step. In total, 1000 emotion images with tags were downloaded from Flickr and defined as  $I_i$ .  $P^{I_i}$  is defined as a set that counts the number of eight basic emotion words contained in  $I_i$ .  $P^{I_i}$  is the set of the number of basic emotion words in which the  $i$ th image was classified through peer evaluation. The probability distribution of the basic emotion words included in the  $i$ th image is  $Y_i$  and is calculated as Equation (2).

$$Y_i = \text{softmax}(P^{I_i}) = \frac{e^{b_k^{I_i}}}{\sum_{u=1}^8 e^{b_u^{I_i}}}, \quad k \in (1, 2, \dots, 8) \quad (2)$$

Finally, to select the emotion images with tags associated with the user's selected lyrics, the representative emotion images selector finds  $\hat{Y}_i$ , which has a probability distribution that is most similar to that of  $U$ , included in  $U^V$ ,  $U^C$ . The index  $i$  of  $\hat{Y}_i$  where the difference in the probability distribution is the minimum, as defined in Equation (3).

$$\hat{i} = \underset{i}{\operatorname{argmin}} (U - V_i)^2, \quad i \in (1, 2, \dots, 1000) \quad (3)$$

The representative emotion images style transfer transfers the styles of  $(n + m)$  emotion images with tags derived from the representative emotion images selector through a CNN model-based style transfer algorithm to the stage background image. Style is features such as color, saturation, brightness, contrast, stroke, and texture. Figure 5 shows the method of outputting an image with a transferred style through a CNN that extracts the image features of content related to the stage background image and a CNN that extracts

the features of style from emotion images with tags. The CNN model normalized the weight of the network using the VGG-16 network, and average pooling was used instead of max pooling. Style characteristics are based on the Gram matrix, ignoring spatial information, and extracting features such as texture and color. Since the correlation of the feature maps of multiple layers, not a single layer, is viewed at the same time, static information, not layout information that the image has globally, is obtained in consideration of multiple scales.

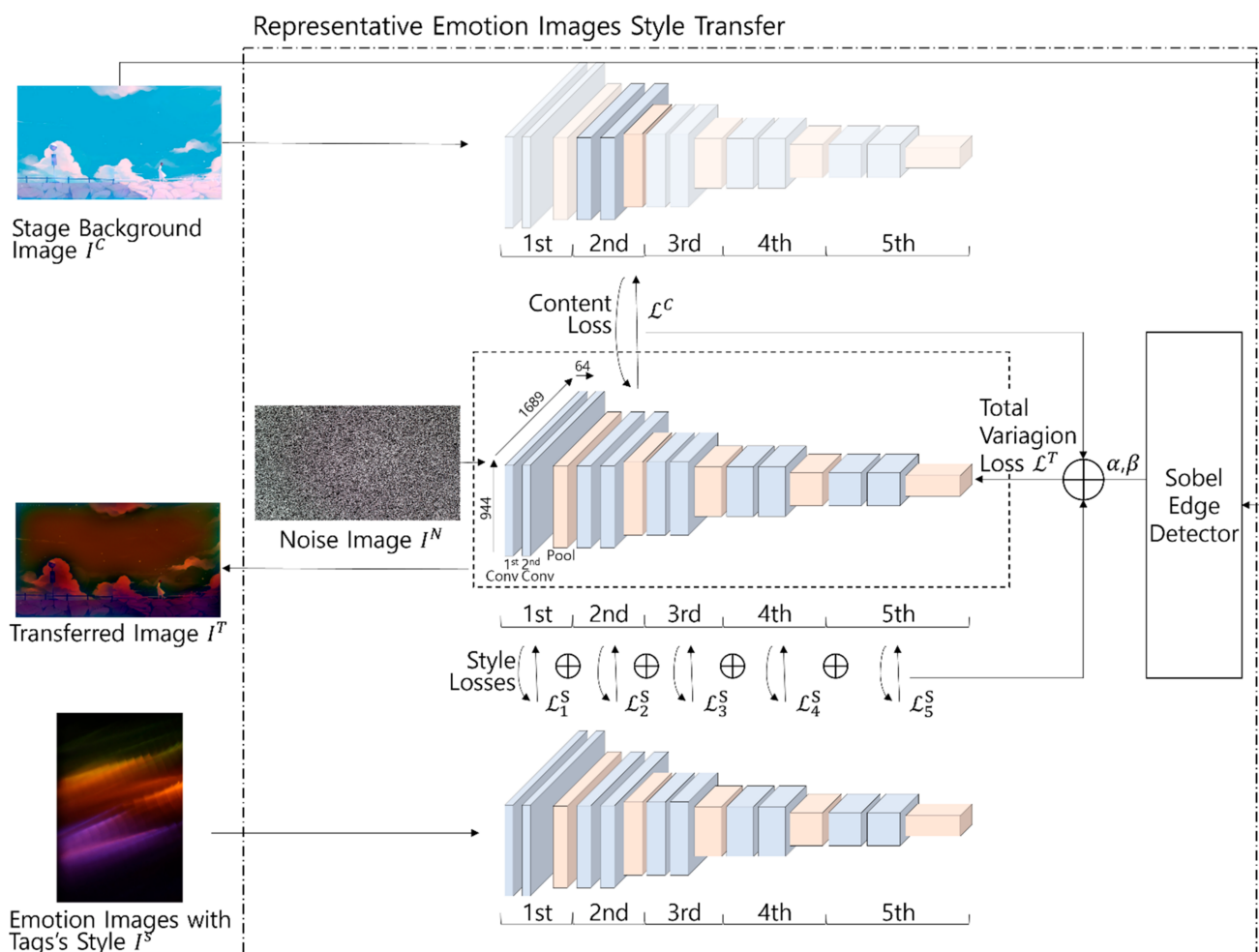


Figure 5. Representative emotion image style transfer process.

Style transfer [9–11] is applied to the representative emotion images style transfer as shown below. There are three types of input images: Stage background image related to the selected lyrics is defined as content image  $I^C$ ,  $I_i$  selected from the representative emotion images selector is defined as style image  $I^S$ , and noise image is defined as noise image  $I^N$ . A noise image is a random variation of brightness or color information in images. This paper synthesizes content of  $I^C$  and style of  $I^S$  on  $I^N$ . The CNN model is composed of a total of five blocks,  $B_1, \dots, B_5$  and one block  $B$  consists of two convolution layers and one pooling layer. After going through one block, each content feature and style feature are extracted. When the input is  $I^C$ , the output from the second block is defined as the content features, and when the input is  $I^S$ , the output at each block is defined as the style features. Content features should have location information of objects included in  $I^C$  and edge information of the object, and the style feature is the correlation of feature maps. The content features of  $I^C$  and the style features of  $I^S$  jointly minimize the distance from the

features of  $I^N$ . Content loss  $\mathcal{L}^C$  is calculated by comparing the content of features of  $I^C$  with  $I^N$  as in Equation (4).

$$\mathcal{L}^C = \left( B_2(I^N) - B_2(I^C) \right)^2 \quad (4)$$

$I^C$  is feed forward through the network. Style loss  $\mathcal{L}^S$ , is calculated by comparing the style features of  $I^S$  with  $I^N$  as in Equation (5).

$$\mathcal{L}^S = \sum_{n=1}^5 \left( B_n(I^N) - B_n(I^S) \right)^2 \quad (5)$$

The pixel-level information disappears as the layer deepens, but the semantic information of  $I^C$  remains the same. Style features should be independent of spatial features. Low-level convolution layers represent low-level features such as edges. This feature maintains a higher resolution. The deeper the layer, the more difficult it is to visualize and interpret features such as edges because they are not directly connected to  $I^C$ . High-level convolution layers capture semantic and less granular spatial information. Style features can get information that considers multiple magnifications of the image globally. However, artifacts occur while transferring  $I^C$  into styles. This implies that  $I^N$ , which is output through the model, loses content information, including the edge information of the objects in  $I^C$ . The deformed error value of the image should be minimized with the transferred style. Edge information of  $I^C$  is recovered through the sobel edge detector [21]. The sobel edge detector is used to reduce the generation of artifacts without losing content features, and then  $\alpha$ ,  $\beta$  is calculated.  $\alpha$  controls the preservation of the content image, and  $\beta$  controls the preservation of the style image. It detects the content features of  $I^C$ , making the content features of  $I^T$  even stronger. The content feature difference between  $I^C$  and  $I^T$  is defined as  $\mathcal{L}^V$  and optimization is performed. In 2D images, sobel edge detection is performed in two directions, vertical and horizontal. The total variation loss  $\mathcal{L}^T$  is calculated based on  $\alpha$ ,  $\beta$ ,  $\mathcal{L}^C$ ,  $\mathcal{L}^S$  as in Equation (6). The noise image  $I^N$  is updated through back propagation based on the total variation loss  $\mathcal{L}^T$ .

$$\mathcal{L}^T = \alpha \mathcal{L}^C + \beta \mathcal{L}^S \quad (6)$$

Optimization proceeds with the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm [22] to find the minimum of  $\mathcal{L}^T$ .

#### 4. Experiments

The probability distribution accuracy verification experiment and the style variation quality verification experiment were performed. It is important that the style of the stage background image is well transformed according to the distribution of emotion words included in the lyrics. This paper verified whether the styles of the stage background image change according to the probability distribution of emotion words included in each verse and chorus and verified the CNN-based style transfer performance.

##### 4.1. Dataset and Experimental Environment

The datasets used to verify the proposed method are the NRC Word-Emotion Association Lexicon (Emolex) and images from Unsplash. The Emolex dataset is a list of English words and their associations with eight basic emotion words (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were performed manually. It includes 6475 English words, and 281 English words were used in the experiment. Table 3 presents the number and distribution of emotion words for each English word to facilitate the use of the Emolex dataset in this experiment. Unsplash is a high-quality open-access image dataset that can be used for further research on machine learning, image quality and search engines. We downloaded 1000 abstract images from Unsplash, and seven colleagues classified them into anger, disgust, fear, sadness, amusement, awe, content, and excitement. Table 3 presents the classified results.

Amusement is a compound emotion of anger and joy, awe is of fear and surprise, content is of joy and trust and excitement is of surprise and joy. The user's selected lyrics used in the experiment was "Forgotten heroes" as shown in Figure 6. The Figure 6 is input to the experiment. The experiments included Windows 10, Intel i7-7700, Nvidia Titan RTX 24 GB graphics card and DDR4 40 GB RAM. The proposed system was developed using Python, and the CNN model was implemented using a deep learning library called Tensorflow.

**Table 3.** Emotion words distribution of Emolex.

	Anger	Disgust	Fear	Sadness	Anticipation	Joy	Surprise	Trust
Quantity	3428 (16%)	3414 (15%)	3572 (17%)	3449 (15%)	2312 (6%)	2325 (10%)	2625 (10%)	2692 (11%)

User's selected lyrics – "Forgotten heroes"

Her alarm goes off And she gets up to watch the morning news Doesn't work no more But tells a lot of stories 'bout her youth Drinks more lately And got pills in many different colors too Morning light is showing She moves the chair to look out at her view But a shop was built right across the street And it stands were the sunrise used to be In the afternoons on the couch to read Goes through old pictures and memories Our heroes have been forgotten Our heroes so brave and bold Our heroes have been forgotten Our heroes oh they got old Our heroes have been forgotten Our heroes so brave and bold Our heroes have been forgotten Our heroes oh they got old Our heroes oh they got old Our heroes oh they got old Smiles they fade because Her daughter only visits once a month Since she got a family of her own It's kept the two apart Used to have so many visitors But now the only one Is the nurse that helps her Move the chair to look out at the sun But a shop was built right across the street And it stands were the sunrise used to be In the afternoons on the couch to read Goes through old pictures and memories Our heroes have been forgotten Our heroes so brave and bold Our heroes have been forgotten Our heroes oh they got old Our heroes have been forgotten Our heroes so brave and bold Our heroes have been forgotten Our heroes oh they got old People don't stay the same you know I just hope their stories will still be told

**Figure 6.** User's selected lyrics.

#### 4.2. Experiment Results

Figure 7 shows the result of extracting verses and choruses from Figure 6 using the verse/chorus extractor. Figure 6 consists of 44 sentences, and each word in each sentence was compared to all the words in entire sentence. A total of 12 consecutive sentences that were repeated twice were extracted as chorus and 20 non-repeated sentences were extracted as verses. Since 44 sentences should be compared with emotion words, the sentences are split into multiple words.

Using the basic emotion words for probability distribution of the emotion word extractor, we compared the emotion words in the lyrics to those in Emolex, as shown in Figure 8. When the words matched, the words in the lyrics were replaced with the basic emotion words. As shown in Figure 8, a total of seven emotion words (alarm, watch, youth, lately, pill, different, and show) were matched in Verse 1; a total of eight emotion words (build, right, stand, couch, hero, old, forgot and bold) in the chorus; a total of five emotion words (hero, old, smile, fade and visit) in Verse 2; and a total of two emotion words (know and hope) in Verse 3. The emotion words were matched a total of 199 times in the song lyrics including duplicates, and the searched emotion words were replaced with the Emolex-based basic emotion words. The probability distributions of the emotion words extractor count the total number of replaced basic emotion words and calculate the probability distribution of each basic emotion word for each verse and chorus.

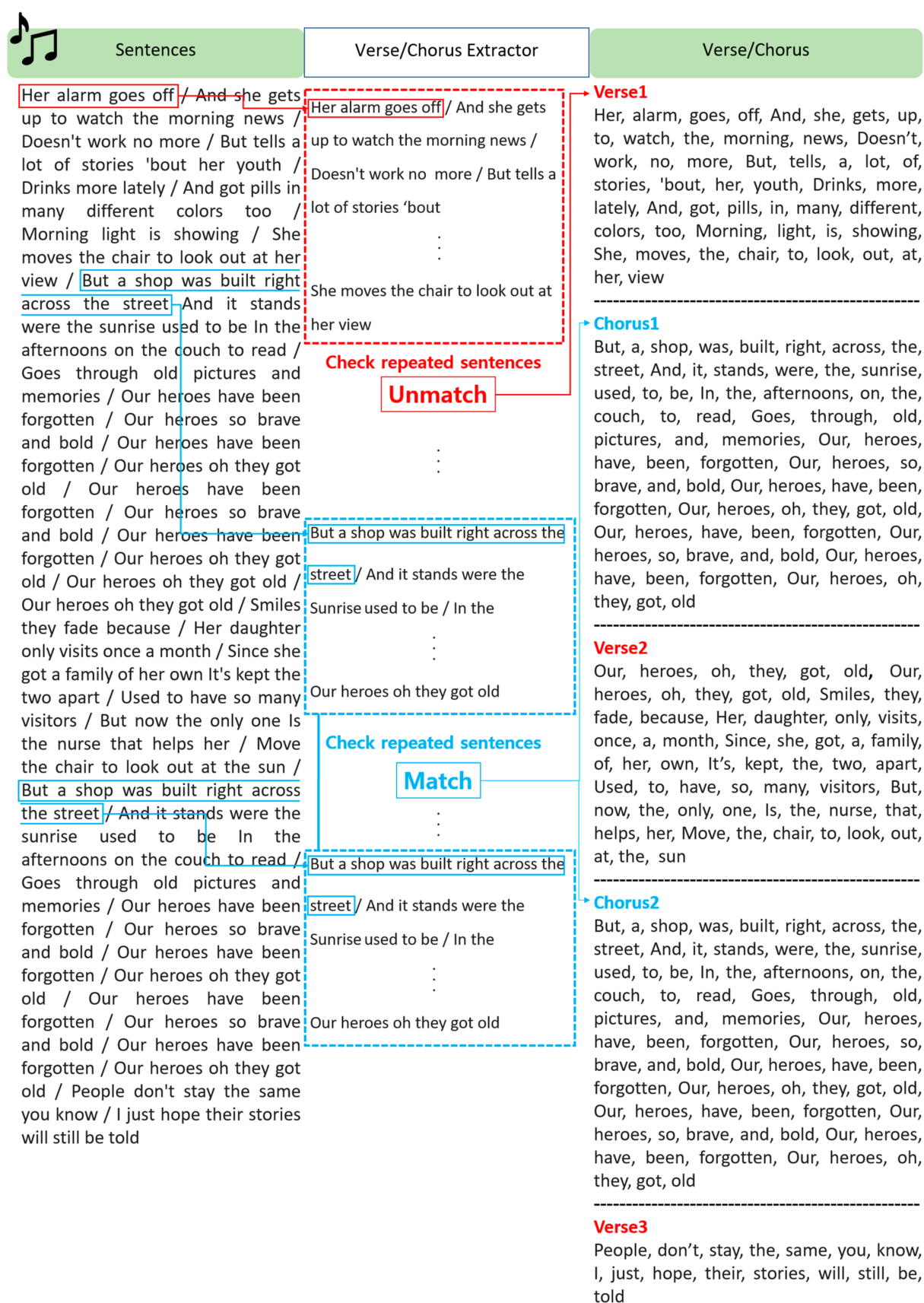


Figure 7. Verse/chorus extractor.

Verses/Choruses	The Basic Emotion Words for Probability Distribution of Emotion Word Extractor	Probability Distributions of Emotion Words
<b>Verse1</b> Her, <b>alarm</b> goes, off, And, she, gets, up, to, watch, the, morning, news, Doesn't, work, no, more, But, tells, a, lot, of, stories, 'bout, her, youth, Drinks, more, lately, And, got, pills, in, many, different, colors, too, Morning, light, is, showing, She, moves, the, chair, to, look, out, at, her, view	<b>Alarm</b> → Fear, Surprise <b>Watch</b> → Anticip, Fear <b>Youth</b> → Anger, Anticip, Fear, Joy, Surprise <b>lately</b> → Sadness <b>Pill</b> → Trust <b>Different</b> → Surprise <b>Show</b> → Trust	<b>Verse1</b> Total Emotion Words: 14 Anger : 1 / 14 7.7% Anticip : 2 / 14 15.4% Fear : 3 / 14 23.1% Joy : 1 / 14 7.7% Sadness : 1 / 14 7.7% Surprise : 3 / 14 23.1% Trust : 2 / 14 15.4%
<b>Chorus1</b> But, a, shop, was, <b>built</b> right, across, the, street, And, it, stands, were, the, sunrise, used, to, be, In, the, afternoons, on, the, couch, to, read, Goes, through, old, pictures, and, memories, Our, heroes, have, been, forgotten, Our, heroes, so, brave, and, bold, Our, heroes, have, been, forgotten, Our, heroes, oh, they, got, old, Our, heroes, have, been, forgotten, Our, heroes, so, brave, and, bold, Our, heroes, have, been, forgotten, Our, heroes, oh, they, got, old	<b>Build</b> → Anticip, Joy, Surprise, Trust <b>Right</b> → Anticip, Joy, Surprise, Trust <b>Stand</b> → Anticip, Joy, Surprise, Trust <b>Couch</b> → Sadness <b>Hero</b> → Anticip, Joy, Surprise, Trust <b>Old</b> → Sadness <b>Forgot</b> → Fear, Sadness <b>Bold</b> → Anticip, Joy, Surprise, Trust	<b>Chorus1</b> Total Emotion Words: 75 Anticip : 15 / 75 20.0% Fear : 5 / 75 6.7% Joy : 15 / 75 20.0% Sadness : 10 / 75 13.3% Surprise : 15 / 75 20.0% Trust : 15 / 75 20.0%
<b>Verse2</b> Our, heroes, oh, they, got, old, Our, heroes, oh, they, got, old, Smiles, they, fade, because, Her, daughter, only, visits, once, a, month, Since, she, got, a, family, of, her, own, It's, kept, the, two, apart, Used, to, have, so, many, visitors, But, now, the, only, one, Is, the, nurse, that, helps, her, Move, the, chair, to, look, out, at, the, sun	<b>Hero</b> → Anticip, Joy, Surprise, Trust <b>Old</b> → Sadness <b>Smile</b> → Joy, Surprise, Trust <b>Fade</b> → Anger, Disgust, Fear, Sadness <b>Visit</b> → Anticip, Joy, Surprise, Trust	<b>Verse2</b> Total Emotion Words: 27 Anger : 1 / 27 3.7% Anticip : 4 / 27 14.8% Disgust : 1 / 27 3.7% Fear : 1 / 27 3.7% Joy : 6 / 27 22.2% Sadness : 3 / 27 11.1% Surprise : 5 / 27 18.5% Trust : 6 / 27 22.2%
<b>Chorus2</b> But, a, shop, was, built, right, across, the, street, And, it, stands, were, the, sunrise, used, to, be, In, the, afternoons, on, the, couch, to, read, Goes, through, old, pictures, and, memories, Our, heroes, have, been, forgotten, Our, heroes, so, brave, and, bold, Our, heroes, have, been, forgotten, Our, heroes, oh, they, got, old, Our, heroes, have, been, forgotten, Our, heroes, so, brave, and, bold, Our, heroes, have, been, forgotten, Our, heroes, oh, they, got, old	<b>Build</b> → Anticip, Joy, Surprise, Trust <b>Right</b> → Anticip, Joy, Surprise, Trust <b>Stand</b> → Anticip, Joy, Surprise, Trust <b>Couch</b> → Sadness <b>Hero</b> → Anticip, Joy, Surprise, Trust <b>Old</b> → Sadness <b>Forgot</b> → Fear, Sadness <b>Bold</b> → Anticip, Joy, Surprise, Trust	<b>Chorus2</b> Total Emotion Words: 75 Anticip : 15 / 75 20.0% Fear : 5 / 75 6.7% Joy : 15 / 75 20.0% Sadness : 10 / 75 13.3% Surprise : 15 / 75 20.0% Trust : 15 / 75 20.0%
<b>Verse3</b> People, don't, stay, the, same, you, know, I, just, hope, their, stories, will, still, be, told	<b>Know</b> → Anticip, Joy, Surprise, Trust <b>Hope</b> → Anticip, Joy, Surprise, Trust	<b>Verse3</b> Total Emotion Words: 8 Anticip : 2 / 8 25.0% Joy : 2 / 8 25.0% Surprise : 2 / 8 25.0% Trust : 2 / 8 25.0%

Figure 8. The basic emotion words for probability distribution emotion word extractor input/output.

The representative emotion image selector compares the similarity with the probability distributions of emotion words extracted through step 1 and probability distributions of the emotion images tags and selects the images for each verse and chorus, as shown in Figure 9. The probability distribution that is similar to the corresponding probability distribution is searched in the dataset of emotion images with tags. Verse 1 has the most similar probability distribution to the probability distribution of (A) and Chorus 1 has the most similar probability distribution to the probability distribution of (B).

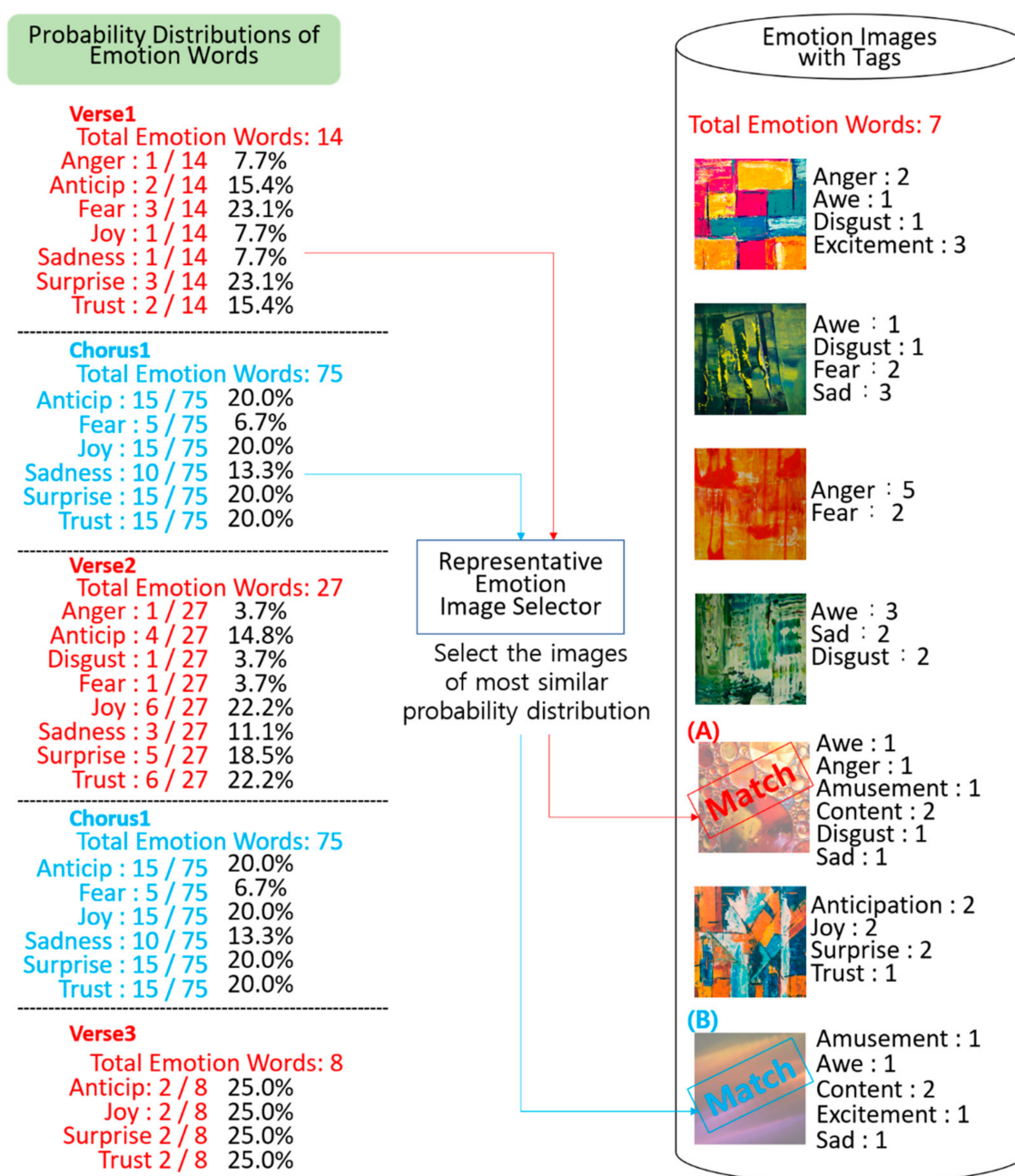


Figure 9. Representative emotion images selector input/output.

However, the disadvantage of this style transfer method is that many high-frequency artifacts occur. The sobel edge detector extracts the edge features of content in the horizontal and the vertical directions because the image is 2 dimension, and edge features of content strengthen. In Figure 10a,b, the edge features of the content are extracted, and it maintains the content edge information well. Figure 10c,d show the high-frequency composition of the image to which the style is applied, but the content edge information is lost as the style is transferred. Figure 11a,b maintain the content feature even when the style is transferred by strengthening the edge feature through the sobel edge detector. As a result, representative emotion image style transfer was output as Table 4, optimization was performed by minimizing style loss and content loss. Figure 12 shows that the total variation loss is minimized from 1.0777 to 0.1597.

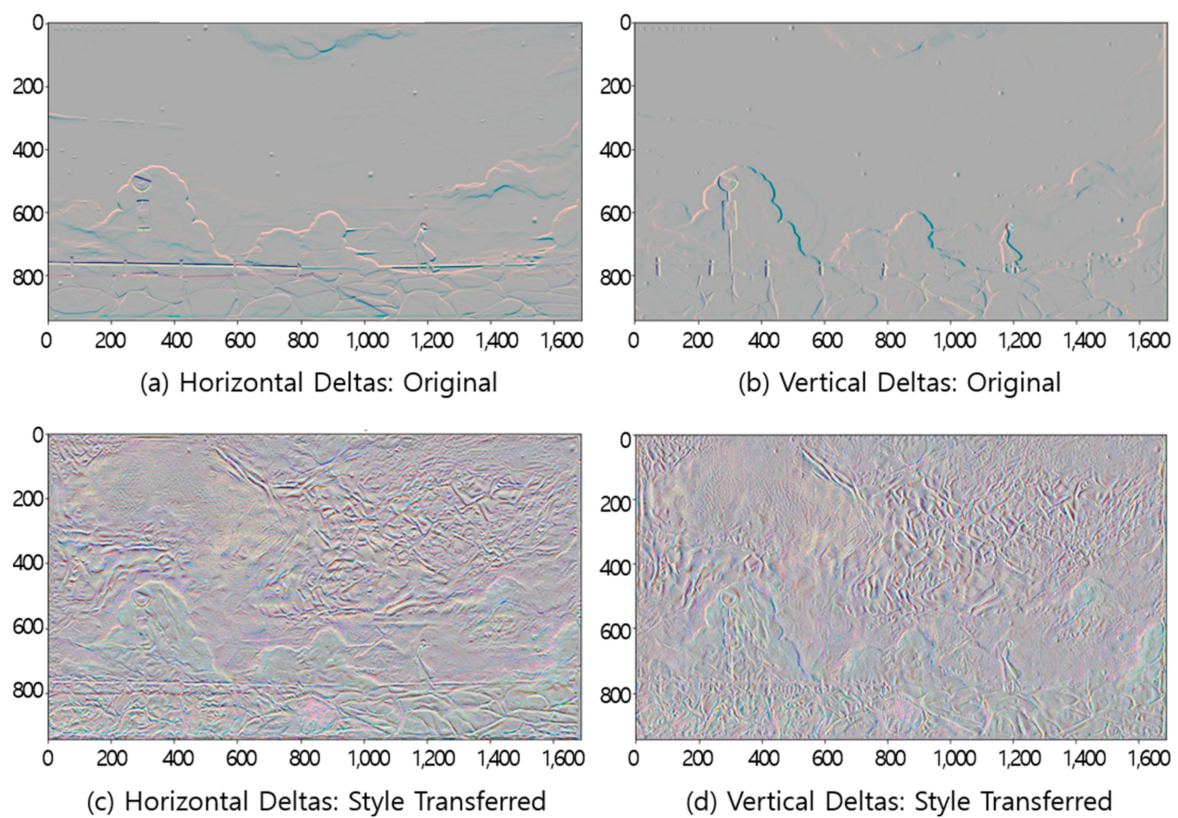


Figure 10. Results of losing content features problem.

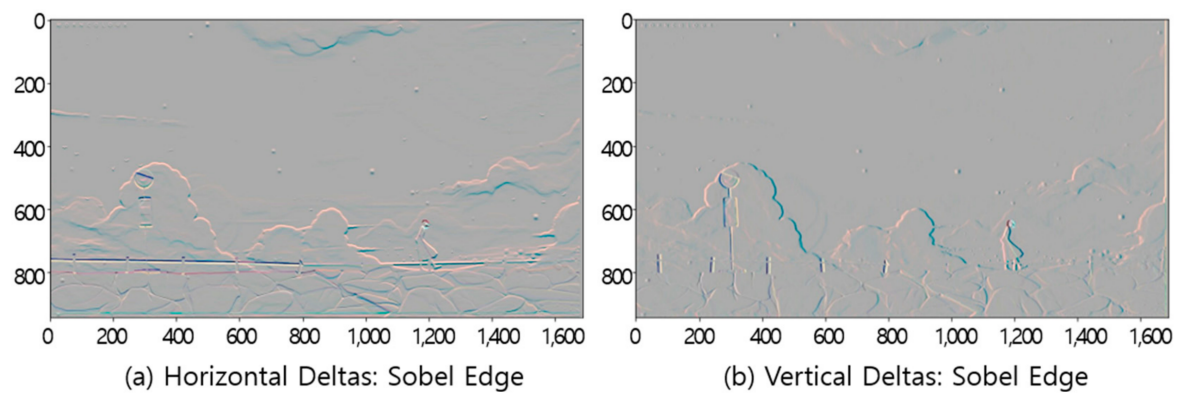
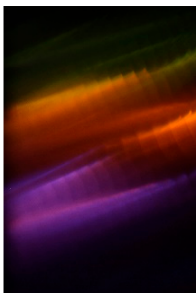


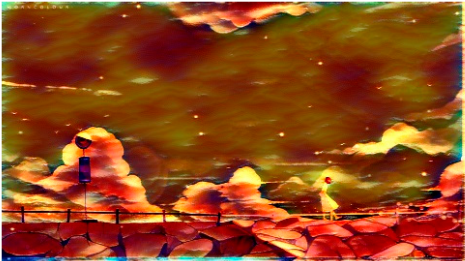


Figure 11. Results of preserving the content features from sobel edge detector.

**Table 4.** Comparison of edges.

Verse 1	Chorus 1
 <p>           Anticip: 0.2            Fear: 0.067            Joy: 0.2            Sadness: 0.133            Surprise: 0.2            Trust: 0.2         </p>	 <p>           Anticip: 0.077            Fear: 0.154            Joy: 0.231            Sadness: 0.077            Surprise: 0.231            Trust: 0.154         </p>
	

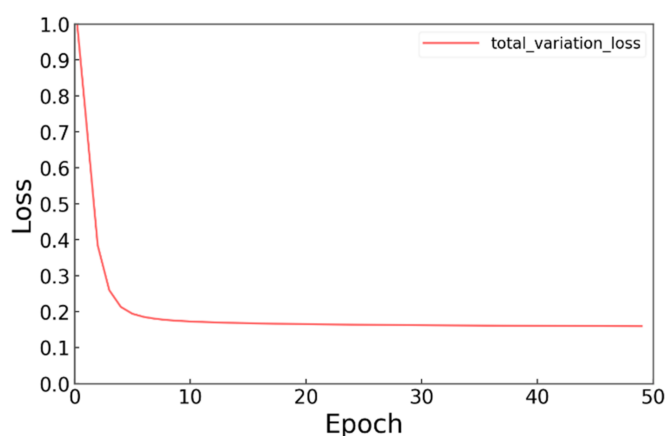




**Figure 12.** Total variation loss.

Table 5 shows the result of comparing the histogram distributions using compareHist function. The styles of images (a), (b) with similar probability distributions of emotion words and an image (c) with a different probability distribution of emotion words were transferred to the stage background image, and the similarity of the images was compared. When the distributions of the pixels of images are similar, the similarity of the images is high, and vice versa. The similarity of images is compared using the compareHist function. The compareHist function allows comparison of image features such as image contrast, color distribution and brightness. (a), (b) and (c) images are compared with the target image for similarity comparison. HISTCMP\_CORREL (correlation) [23] is a correlation expressed by calculating pixels having the same value and is calculated as in Equation (7). The closer the value is to 1, the more similar the images are.  $H$  is a histogram, and  $N$  is the total number of histogram bins.

$$d(H_1, H_2) = \frac{\sum_J (H_1(J) - \overline{H_1})(H_2(J) - \overline{H_2})}{\sqrt{\sum_J (H_1(J) - \overline{H_1})^2} \sqrt{\sum_J (H_2(J) - \overline{H_2})^2}} \text{ where, } \overline{H_k} = \frac{1}{N} \sum_T H_k(T) \quad (7)$$

**Table 5.** Result of histogram comparison using compareHist function.

				
	Target image	Image with a similar distribution (a)	Image with a similar distribution (b)	Image with a different distribution (c)
Emotion probability distribution	Amusement: 0.14 Awe: 0.14 Content: 0.29 Excitement: 0.29 Sad: 0.14	Amusement: 0.14 Awe: 0.29 Content: 0.29 Excitement: 0.14 Sad: 0.14	Amusement: 0.29 Awe: 0.14 Content: 0.29 Excitement: 0.29	Anger: 1.00
HISTCMP_CORREL	1.00	0.12	0.22	0.1
HISTCMP_CHISQR	0.00	7665.83	7443.83	17,372.30
HISTCMP_INTERSECT	1.00	0.38	0.28	0.05
HISTCMP_BHATTACHARYYA	0.00	0.62	0.62	0.92

HISTCMP\_CHISQR (Chi-squared distribution) [23] is the distribution of the spread of pixel values. It is calculated as in Equation (8) and the closer it is to 0, the more similar the images are.

$$d(H_1, H_2) = \sum_J \frac{(H_1(J) - H_2(J))^2}{H_1(J)} \quad (8)$$

HISTCMP\_INTERSECT (intersection) [23] computes the similarity of two discrete probability distributions, as in Equation (9), using the possible values of the intersection between 0 and 1. The closer to 1, the more similar the images are.

$$d(H_1, H_2) = \sum_J \min(H_1(J), H_2(J)) \quad (9)$$

HISTCMP\_BHATTACHARYYA [24] calculates the degree of overlap of two probability distributions as in Equation (10). The closer to 0, the more similar images are.

$$d(H_1, H_2) = \sqrt{1 - \frac{1}{\sqrt{H_1 H_2} N^2} \sum_J \sqrt{H_1(J) \cdot H_2(J)}} \quad (10)$$

Table 6 shows histogram graph according to RGB distribution, hue, and value. The horizontal axis of the graph represents the change in color tone from 0 to 255, with the left side representing the dark area and the right side representing the bright area. The vertical axis of the graph represents the size of the area captured in each horizontal area, that is, the total number of pixels. This is the number of pixels in an image over a range of 256 pixel values.

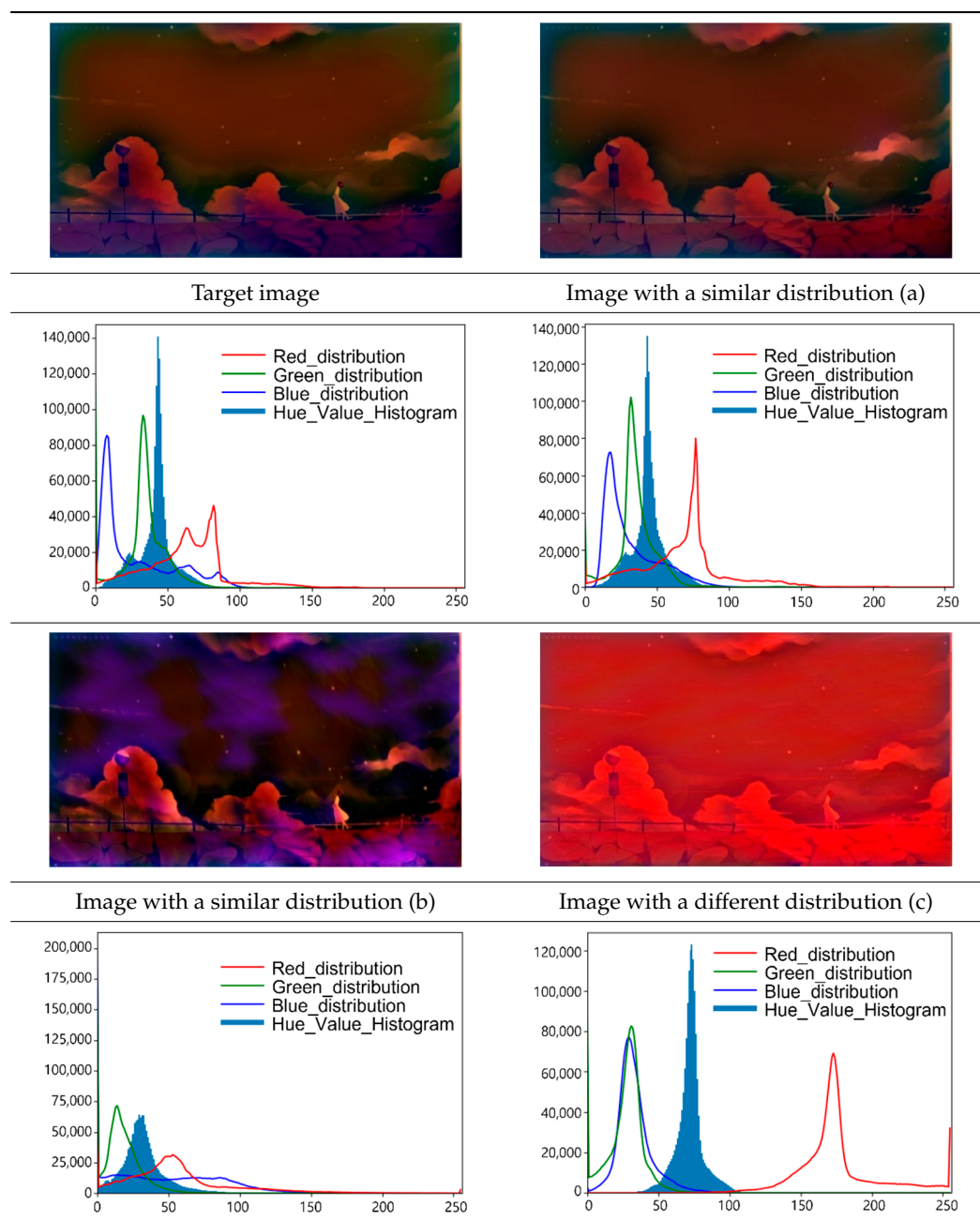

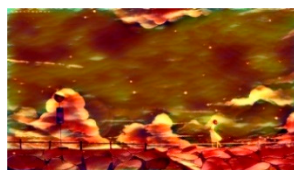









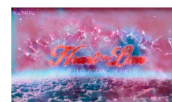
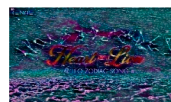
**Table 6.** Results of histogram graph according to RGB distribution, hue, and value.

Table 7 shows the results of inputting various lyrics. By inputting the song lyrics of “Meant to be this way”, “Sax is my cardio” and “Heart of lion (Leo)”, the stage background images were transformed according to the proposed method for each verse and chorus.

Table 7. Results of inputting various lyrics.

Lyrics	Verse 1	Chorus	Verse 2	
Forgotten hero				
Meant to be this way				
Sax is my cardio				
Heart of a lion (Leo)	Verse 1	Chorus	Verse 2	Verse 3
				

The song lyrics “Meant to be this way” is consisted of two verses and two choruses, and a total of 12 emotion words were extracted. In this song’s lyrics, 4 emotion words out of 14 sentences in verse, 5 emotion words out of 14 sentences in chorus, and 3 emotion words out of 14 sentences in verse 2 were extracted. The song lyrics to “Sax is my cardio” is consisted of two verses and two choruses, 28 emotion words were extracted. In this song’s lyrics, 14 emotion words out of 12 sentences in verse 1, 7 emotion words out of 8 sentences in chorus, and 7 emotion words out of 12 sentences were extracted in verse 2. The song lyrics “Heart of a lion (Leo)” is consisted of three verses and two choruses, and a total of 32 emotion words were extracted. In this song’s lyrics, 5 emotion words out of 8 sentences in verse 1, 10 emotion words out of 13 sentences in chorus, 3 emotion words out of 8 sentences in verse 2 and 9 emotion words out of 8 sentences in verse 3 were extracted. The styles were transferred by selecting the images with the most similar probability distributions for each verse and chorus through the emotion words extracted from each song lyrics. We confirmed through the results in Table 7 that the styles are well transformed even from complex stage background images.

## 5. Conclusions

This paper proposed a method to transfer stage background images into styles based on the emotion words contained in each verse and chorus from lyrics selected by a user. First, multiple verses and choruses were derived from the lyrics, one at a time, and compared with the emotion word dictionary to extract the emotion words included in each verse and chorus. Next, the image with the most similar probability distribution to the corresponding probability distribution was selected based on the probability distribution of emotion words included in the lyrics, and the styles were transferred to the stage background image for each verse and chorus. In the experiment, the performance of the style transfer was verified, and the probability distribution of the emotion words in the transformed stage background image was verified as similar to the probability distributions of the song lyrics. Experimental results showed that the proposed method reduced total

variation loss from 1.0777 to 0.1597. This result shows that the style transferred image is close to edge information about the content of the input image, and the style is close to the target style image. In addition, stage background image and images of transferred styles with similar emotion words probability distributions were 38% similar, and stage background image and image of transferred styles with completely different probability distributions were 8% similar.

Due to the limitations of lexicon-based approaches, several aspects related to the design of relevant emotion analysis models need to design a model in future works. The input of the models that extract emotions by considering full sentences is sentence, but the lyrics do not follow the complete sentence structure. It is difficult to use each structure as an input to the previous models. In the case of a full sentence, there is a limit to the accuracy because there is uncertainty of specifying all emotions corresponding to the words of each sentence. Therefore, in this paper, limited emotion words were selected and utilized.

**Author Contributions:** Conceptualization, H.Y., S.L. and Y.S.; methodology, H.Y., S.L. and Y.S.; software, H.Y., S.L. and Y.S.; validation, H.Y., S.L. and Y.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Ministry of Science, ICT (MSIT), Korea, under the High-Potential Individuals Global Training Program, grant number 2019-0-01585 and 2020-0-01576 supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and the APC was funded by 2019-0-01585 and 2020-0-01576.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** This manuscript is one of the results by the project funded by the Ministry of Science, ICT (MSIT), Korea under the High-Potential Individuals Global Training Program, grant number 2019-0-01585 and 2020-0-01576 supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Li, S.; Jang, S.; Sung, Y. Melody Extraction and Encoding Method for Generating Healthcare Music Automatically. *Electronics* **2019**, *8*, 1250. [\[CrossRef\]](#)
2. Li, S.; Jang, S.; Sung, Y. Automatic Melody Composition Using Enhanced GAN. *Mathematics* **2019**, *7*, 883. [\[CrossRef\]](#)
3. Wen, J.; She, J.; Li, X.; Mao, H. Visual Background Recommendation for Dance Performances Using Deep Matrix Factorization. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2018**, *14*, 1–19. [\[CrossRef\]](#)
4. Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; He, X. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
5. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1947–1962. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Hu, X.; Downie, J.S.; Ehmann, A.F. Lyric Text Mining in Music Mood Classification. In Proceedings of the 10th International Society for Music Information Retrieval (ISMIR), Kobe, Japan, 26–30 October 2009.
7. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
8. Hori, G. Color Extraction from Lyrics. In Proceedings of the 2019 4th International Conference on Automation, Control and Robotics Engineering (CACRE), Shenzhen, China, 19–21 July 2019.
9. Isola, P.; Zhu, J.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
10. Sung, Y.; Jin, Y.; Kwak, J.; Lee, S.; Cho, K. Advanced Camera Image Cropping Approach for CNN-Based End-to-End Controls on Sustainable Computing. *Sustainability* **2018**, *10*, 816. [\[CrossRef\]](#)
11. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image Style Transfer Using Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
12. Zhao, S.; Gao, Y.; Jiang, X.; Yao, H.; Chua, T.; Sun, X. Exploring Principles-of-Art Features for Image Emotion Recognition. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014.

13. Machajdik, J.; Hanbury, A. Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th ACM International Conference on Multimedia, Florence, Italy, 25–29 October 2010.
14. Han, E.; Cha, H. Extraction of Critical Low-Level Image Features for Effective Emotion Analysis. *Inst. Control. Robot. Syst.* **2019**, *25*, 319–326. [\[CrossRef\]](#)
15. Wei, Z.; Zhang, J.; Lin, Z.; Lee, J.; Balasubramanian, N.; Hoai, M.; Samaras, D. Learning Visual Emotion Representations from Web Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020.
16. Yang, D.; Lee, W. Music Emotion Identification from Lyrics. In Proceedings of the IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 14–16 December 2009.
17. Zhao, S.; Yao, H.; Gao, Y.; Ding, G.; Chua, T. Predicting Personalized Image Emotion Perceptions in Social Networks. *IEEE Trans. Affect. Comput.* **2018**, *9*, 526–540. [\[CrossRef\]](#)
18. Lee, J.; Lim, H.; Kim, H. Similarity Evaluation of Popular Music based on Emotion and Structure of Lyrics. *KIISE Trans. Comput. Pract.* **2016**, *22*, 479–487. [\[CrossRef\]](#)
19. NRC Word-Emotion Association Lexicon. Available online: <http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm> (accessed on 31 December 2020).
20. Mohammad, S.M.; Turney, P.D. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In Proceedings of the Computational Approaches to Analysis and Generation of Emotion in Text(CAAGET), Los Angeles, CA, USA, 13–19 June 2010.
21. Gao, W.; Zhang, X.; Yang, L.; Liu, H. An improved Sobel edge detection. In Proceedings of the 2010 3rd International Conference on Computer Science and Information Technology, Chengdu, China, 9–11 July 2010; Volume 5, pp. 67–71.
22. Zhu, C.; Richard, H.B.; Lu, P. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw. (TOMS)* **1997**, *23*, 550–560. [\[CrossRef\]](#)
23. Ogul, H.; Celik, N. A Web Application for Content based Geographic Image Retrieval. In Proceedings of the 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, 15–18 May 2017; pp. 1–4.
24. Choi, E.; Lee, C. Feature extraction based on the Bhattacharyya distance. In *Pattern Recognition*; Elsevier: Amsterdam, The Netherlands, 2003; Volume 36, pp. 1703–1709.