# Generation of Photo Slideshow with Song based on Closeness between Concept of Lyrics and That of Images

Mei HASHIMOTO* and Michiharu NIIMI†
* Graduate School of Computer Science and Systems Engineering
Kyushu Institute of Technology, Iizuka, Japan
E-mail: hashimoto.mei582@mail.kyutech.jp
† Faculty of Computer Science and Systems Engineering
Kyushu Institute of Technology, Iizuka, Japan
E-mail: niimi@ai.kyutech.ac.jp

***Abstract*—This paper proposes a method that allows users to easily convert a large number of still images into movies by displaying photos in sync with memories or favorite songs, which offers a new media viewing method for users to look back at the vast number of photos in their photo folders. We assume here that users select a song and have images stored in local PC. One method of synchronization involves selecting images based on elapsed time of song. However, since lyrics convey meaning, it can be better to display the images that align with both lyrics' meaning and images' meaning matched. In order to do this, we need the space which shares the concept of both lyrics and images, and select colors as the concept. Displayed images are determined based on the distance between points that are mapped from words and images to color space. Specifically, the color concepts from words are extracted using the Color Image Scale, and that from images are determined as the most frequent colors by cluster analysis. Experimental results show that it is possible to easily create movies where images with colors that match the mood of the lyrics are displayed.**

## I. Introduction

In recent years, the household penetration rate of mobile devices has exceeded 97%, and their evolution has been remarkable. In particular, smartphones have rapidly spread and are equipped with increasingly advanced features. This enables us to capture high-quality photos of everyday moments as memories and easily share them with friends. The emergence of large-capacity storage and cloud services allows for worry-free photo-taking without concern for storage limitations. Therefore, the number of photos in photo folders continues to increase. However, we rarely look back at these accumulated photos.

Additionally, the use of movie-sharing sites and social media platforms such as YouTube, TikTok, and Instagram has increased rapidly, which provides excellent ways for easily sharing information with other users. Currently, it is common to see users who post dance movies synchronized with popular songs or musics or lyric movies that align photos with the meaning and timing of the lyrics of memorable songs. Combining songs and visuals that match in impression can enhance and emphasize each other's impact. Many users may be interested in content that synchronizes with music and lyrics with natural. Therefore, the ability to easily create visual content that matches music is considered highly valuable. Such synchronized visual content includes still images, therefore, it is one possible approach to utilize the accumulated photos.

There are several tools that users manually synchronize music and photos to create movies. For instance, in Clipchamp of Microsoft, users select BGM from available options, choose image files, and manually set/edit display timing and sequence to create movies. Additionally, there are automated movie creation services such as iPhone's "For You" and "Google Photos." "For You" enhances enjoyment of photos and movies in the iPhone's standard photo library by automatically organizing them based on facial detection and image analysis. Using its "Memories" feature, it automatically picks photos and movies for specific dates, locations, people, or events and creates slideshow movies combined with BGM. Google Photos features an "Auto Creations" function offering many themes like growth records, annual highlights, and dog movies, which automatically select appropriate photos stored within Google Photos to create 10 to 20-second slideshow movies accompanied by themed BGM. However, these tools do not generate movies synchronized with music lyrics.

Therefore, in this study, we propose a method to automatically display photos that match favorite musics with lyrics. It can be a new viewing of photos method for users to look back at a large collection of photos in their photo folders. We aim to develop a method that can easily generate slideshow movies where images switch in sync with lyrics automatically from a set of still images when given a piece of music.

Recently, there are researches which have focused on selecting images based on lyric information. Ishizaki et al.[1] extract keywords (overall impression words) from lyrics and further derive search queries for images from specific lyric lines. They use these search keywords to select similar images available on the web. They leverage the tags associated with web images. This method assumes scraping lyric information from lyric-sharing sites where time information is not attached

to lyrics, thus not enabling synchronized display of lyrics and images as they appear. Moreover, it relies on tagged images. Because tags are not typically put to photo collections in local folder, this approach is unsuitable for our purpose.

Furthermore, there is a research of concept acquisition extracting "meaning" from different media types such as audio and images. Ohishi et al.[6] propose a new framework for extracting targeted speaker conversations from mixed audio using semantic clues. By mapping semantically related images and audio to a shared embedding space, they effectively extract desired audio content.

For the purpose of automatic synchronization between lyrics and images in our study, it is essential to compute the correlation between lyrics of a song and images in a suitable feature space, which is denoted as shared concept space.

For instance, Niiho et al.[5] construct a comprehensive impression word space using canonical correlation analysis based on image and music feature vectors and their impression words. By mapping an unknown image's feature vector into this space, they recommend a music that closely match its impression from a pre-learned set of music.

There is also research converting extracted features from lyrics into visual information. Midorikawa et al.[3] directly correlate two music features: the average of tempo and the ratio of high-frequency to image saturation and brightness. In addition, Murata et al.[4] convert lyrics into color vectors using the Color Image Scale[2], which is denoted as CIS. This scale comprehensively expresses human-standard sensations from colors by using words that express emotion and sensibility adequately, which are referred to as "impression words", CIS describes the relationship of colors and impression words. Several impression words corresponds to one color. They define music color vectors as the color vectors extracted from lyrics using CIS, then "color changes" are generated for the entire music due to lyrics consisting of many word strings. They recommend similar music based on this similarity metric based on the color changes.

In this paper, as the first step of our research, we select a color space as the shared concept space. Specifically, we extract color information from lyrics using CIS and designate the most frequent color from images as their color information. Based on the similarity of this color information, we determine which images should display. Thus, we compare the similarity between lyrics and images in the color space. Firstly, for given a piece of music, we extract lyric data annotated with time information automatically. Next, from the lyric data, we extract impression words and convert them into colors based on CIS. Using this information and matching it with the most frequent color of images, we select images that match the lyrics.

In Section II, we describe the proposed method, then, in Section III, we presents the experiments conducted with our method and their results, and discuss it. Finally, in Section IV, we summarize and discuss future challenges and prospects.

## II. Proposed Method

### A. System Overview

We have a piece of music with lyrics, that is a song, and a collection of digital photos (Digital Image) taken with digital devices such as smart phones. The music prepared by the user is referred to as $M$, and the collection of digital photos is denoted as $D = \{d_i | i = 1, 2, \ldots, n\}$($D$ consists of $n$ digital photos).

First, keywords are extracted from $M$. These are represented as $kw_j (j = 1, 2, \ldots, m)$(where $m$ keywords are extracted from the music M). The time (singing time) when each $kw_j$ is sung from the beginning of the song is denoted as $st_j$. That is, $kw_m$ is sung after $st_m$ elapsed time from the beginning of the song.

If we denote the function that maps specific keywords to their concept space as $C_{KeyWord}$, and the function that maps image data to their concept space as $C_{DigitalImage}$, for each $kw_i$,

$$\underset{i}{\text{argmin}} \quad \text{distance}(C_{KeyWord}(kw_j), C_{DigitalImage}(d_i)) \quad (1)$$

we find $i$ that satisfies the above equation. Here, distance($x$,$y$) represents the distance between $x$ and $y$. By performing this process for all $kw_j$, each $kw_j$ is associated with the one image $d_i$ that is closest in the concept space.

Utilizing the information of $st_j$, we can arrange $d_i$ on the time axis. When we make a movie from the set of images with $M$ as the BGM, images that correspond to the concepts of the lyrics would be displayed following lyrics' meanings. This is the image display system that we propose in this study, which aims to synchronize the display of images with the concepts of the lyrics by leveraging the proximity of concepts between the lyrics and the images.

We set the concept for songs, which are actually lyrics, and images, and construct an image display system based on the similarity in the concept space. Fig.1 illustrates the outline of our proposed system. Similarly to Murata et al.[4], the shared concept space, which is the common concept to be extracted from both songs and images, is set as "color," The color concept from the lyrics is generated based on CIS, and the color concept from the images is defined as the most frequent color in the respective image.

Each process is explained in detail below.

*1) $C_{KeyWord}$:* The function $C_{KeyWord}$, which maps specific keywords extracted from music lyrics into a shared concept space, consists of three components: music_spleeter, whisper and music_concept. In music_spleeter, we use a library called spleeter, which can split the sounds of each instrument in the music using machine learning, to separate them into Vocal, Piano, Bass, Drums, and other. Then, vocal audio files are input into the Whisper speech recognition model. As a result of this, we can get lyrics from the song. The lyrics is consisting of word strings which are complete sentence and/or phrase making a certain meaning, and we call it a quasi-sentence. From a lyrics, we can extract multiple quasi-sentences. In addition, the Whisper output the starting

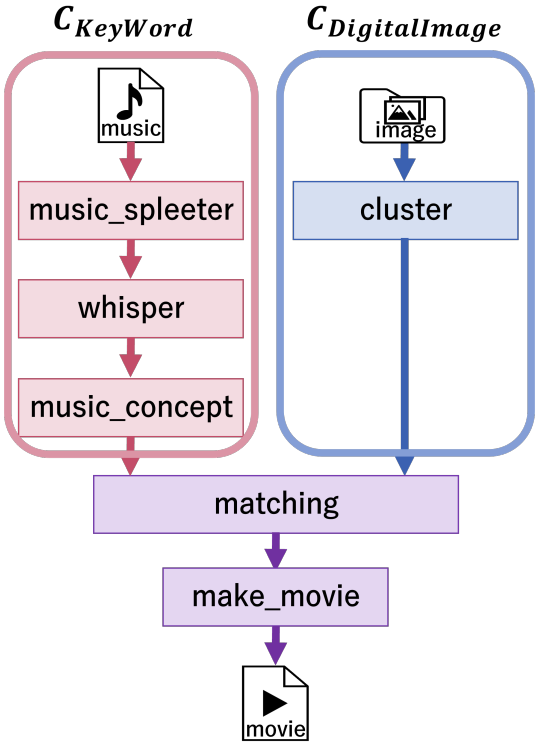



**Fig. 1:** Overview of the Image Display System

time of each quasi-sentence. For each quasi-sentence, we employ morphological analysis with mecab to extract nouns, verbs, and adjectives which are regarded as keywords for the quasi-sentence. Finally, a quasi-sentence is mapped into a impression word of CIS. This is done by the followings in detail. A quasi-sentence includes multiple keywords, and there are multiple impression words we pre-defined to use our system. We calculate a similarity between all keywords and all impression words. Based on this calculation, we extract a pair of keyword and impression word that minimizes the distance in shared concept space, call the keyword of the pair "a representative word" of quasi-sentence.

*2) $C_{DigitalImage}$:* The function $C_{DigitalImage}$ mapping image data to the concept space performs clustering using the K-means method. The number of clusters is set to five, and the cluster center of the cluster with the largest number of data points is extracted as the most frequent color of the image data. This color is then mapped to the concept space to set the concept.

*3) matching:* After extracting the concepts from the lyrics and the images, we link each keyword to the image that minimizes the distance in the shared concept space between the lyrics and the images by following equation (1). This process is performed for all extracted lyric concepts. Using the timestamps of the lyrics, the image data linked to the lyrics are arranged accordingly on the time axis.

*4) make_movie:* Based on the arranged image data, a movie is created by displaying images along with the music $M$. To minimize the discrepancy between the appearance of words and the display of images, fade-in and fade-out image processing is applied. The crossfadein and crossfadeout functions from the MoviePy library are used for the fade-in and fade-out effects. The current image is faded out over one second while the next image is faded in simultaneously. To ensure that images can be properly appreciated and are not difficult to see, images with a display time of three seconds or less will not undergo fade-in and fade-out processing.

**TABLE I:** Used songs

| Melody | Song title/artist | Performance time |
|---|---|---|
| bright song | shinning star / shiho | 4m42s |
| sad song | Diamond Dust-tribute to Frozen- / aruku | 4m18s |
| refreshing song | Success! / TOYro | 1m24s |

## III. Experiment

### A. Experimental condition

In this experiment, we used three free music tracks : bright song, sad song, and refreshing song, as shown in Table I. We also collected 524 images taken by students in my laboratory. Additionally, we selected 27 categories of impression words. The relationship between the impression words and colors in the shared concept space is shown in Table II.

### B. Results

Using the proposed system, we were able to extract concepts from both the music and the images, and create movies using images from the image folder. The results of this experiment are presented for each structure based on the overview of the system in Section II-A. Additionally, the movies created were shared within the member of my laboratory, and evaluated it. We also describe this evaluation.

*1) Song concept extraction:* In music_spleeter, we were able to separate the song into Vocal, Piano, Bass, Drums, and other components. The Vocal audio files were transcribed into lyrics using Whisper, and morphological analysis was performed using mecab, then mapped to the shared concept space. For the multiple keywords obtained from the lyrics, the most similar representative word was mapped to the shared concept space to obtain the concept (color). Table III, Table IV and Table V show the results of the concept extraction for the chorus parts of the three songs used in this experiment.

*2) Image concept extraction:* For 524 images, cluster analysis was performed using the K-means method to extract the most frequent color representing the concept of each image. Table VI shows the images subjected to concept extraction, the most frequent colors obtained through cluster analysis, and the concepts (colors) obtained by mapping these most frequent colors to the shared concept space.

*3) Concept matching:* By following the proposed steps, we were able to obtain the concepts from both quasi-sentences of the song and the images. Using the obtained concepts, we linked the images with quasi-sentences that distributes

**TABLE II:** Impression words and their corresponding colors [2]

| Impression words | Color (R,G,B) |
|---|---|
| cute, innocent, childlike, charming, pretty | (239,143,184) |
| happy, joyful, cheerful, energetic, casual | (255,88,80) |
| splendid, bustling, vivid, flashy | (192,0,112) |
| active, dynamic, bold, exciting, passionate | (0,140,113) |
| dynamic,energetic,intense,forceful | (16,0,32) |
| gorgeous, glamorous, sexy, rich, luxurious | (132,48,143) |
| wild, powerful, tough, robust, manly | (0,117,62) |
| lovely, sweet, fresh | (255,212,200) |
| mild, familiar, gentle, pleasant | (255,213,159) |
| natural, honest, relaxed, simple, unhurried | (198,205,156) |
| graceful, feminine, emotional, delicate, beautiful | (219,149,173) |
| classic, nostalgic, antique, traditional, meticulous | (104,0,31) |
| strong, solid, heavy | (0,60,60) |
| modest, delicate, feminine, elegant | (197,184,199) |
| elegant, humble, dignified | (165,129,145) |
| authentic, dignified, refined, robust | (85,0,53) |
| romantic, fantastical, pale, pure, neat | (225,236,255) |
| juicy, reserved, plain, peaceful | (217,253,255) |
| lively, healthy, fresh, safe | (218,248,109) |
| quiet, subtle, reserved, nonchalant, simple | (124,152,156) |
| smart, stylish, intellectual, calm, subdued | (0,50,117) |
| dandy, gentlemanly, masculine, serious, steady | (0,71,91) |
| formal, noble, sacred, solemn | (107,107,98) |
| brisk, clean, pure, clear, shipshape | (138,176,223) |
| refreshing, youthful, young, sporty | (91,189,206) |
| modern, speedy, innovative, agile, intelligent | (54,119,133) |
| mechanical, precise, rational, artificial | (161,161,148) |

with near distance in shared concept space. We were then able to arrange the appearance times of the lyrics and the corresponding image file names as data on the timeline.

*4) Make movie:* Following the timeline data obtained by the above Section III-B3 (concept matching), we were able to create a movie synchronized with the music by displaying the images in the order aligned with the music as the BGM (Background Music). For the created movie, the keywords from the lyrics, impression words, and the images displayed at those times are shown in Table VII, VIII and IX based on III-B1) Music Concept Extraction.

*5) Subjective evaluation:* The created movie was shared within the member of my laboratory, and evaluated with subjectively. The summarized results are presented below.

- Regarding synchronization between the music and images, there are instances where images aligned well with the concept and mood of the lyrics. But there are also parts where the relationship between lyrics and images is unclear.
- To improve the quality of the movie and clarify the relationship between music and images, it would be beneficial to display lyrics when images corresponding to the concept of the lyrics are shown.
- Although the purpose of this system is different, it my be interesting using AI-generated images based on impression words and lyrics.

**TABLE III:** Concept of song:shinning star

| Keyword | Representative word | Impression word | Color (R,G,B) |
|---|---|---|---|
| wind, sound, confine | sound | mild | (255,213,159) |
| dream, sleep, illusion | dream | cute | (239,143,184) |
| world | world | nature | (198,205,156) |

**TABLE IV:** Concept of song:Diamond Dust-tribute to Frozen-

| Keyword | Representative word | Impression word | Color (R,G,B) |
|---|---|---|---|
| sun, see, scenery, forget, white, diamond | scenery | romantic | (225,236,255) |
| living, proof, give | proof | safety | (218,248,109) |
| beside, stay | beside | antique | (104,0,31) |

## C. Discussion

Using the proposed model, we synchronized image display with music based on the proximity of lyrical and image concepts. We were able to display images that matched the mood of the lyrics in the music. In this study, Despite setting only color as a concept, there were several instances where subjects matching lyrical words were displayed. Specifically, in Table IV, the keyword “scenery” was assigned a light blue concept. When selecting images that matched the concept of the lyrics from the image folder at random, a photo depicting the sky, sea, and a brightly shining white sun was coincidentally chosen. This means that we can select images that perfectly matched the lyrics “day, scenery, forget, white diamond.” However,

**TABLE V:** Concept of song:Success!

| Keyword | Representative word | Impression word | Color (R,G,B) |
|---|---|---|---|
| tomorrow, extend, grasp | tomorrow | gentle | (255,213,159) |
| future, hand | hand | strong | (0,60,60) |
| match, time, suddenly | match | clear | (138,176,223) |

challenges arise when there are prominent subjects in the image or when keywords like "nature" have multiple meanings. This leads to the selection of images that do not align with our intended imagery. For instance, photos containing family, friends, lovers, or pets in the image folder may not match well with lyrical words appearing in the lyrics. Additionally, the concept varies depending on whether it is mapped to the shared concept space as "nature" in general or as "natural" in a specific sense. Therefore, it is necessary to set keywords considering not only individual words but also their context.

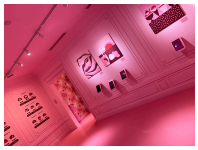
Furthermore, considering insights gleaned from comments on the created movie, potential improvements are discussed below.

Based on the data files storing the conceptual data of the songs and images created in this experiment, it appears that the concepts for both songs and images were appropriately set. However, there are three potential reasons for the unclear relationship between quasi-sentences and images in certain parts of the created movie.

(1) It may be insufficient for extracting image concepts. In this study, the most frequent color in the image is used as the feature for image representation. Therefore, as seen in photos of strawberries and fireworks in Table VI, the extracted color represents the background color rather than the color of the strawberries or fireworks that we want to focus on. This misalignment can lead to attention being drawn away from our intended image and impressions. In the future, incorporating methods that specify attention positions could enable the creation of movies more synchronized with the music.

(2) It may be insufficient for matching the concepts of lyrics and images. In this paper, the system is set to select the image with the closest associated impression word even though the distance is far in the shared concept space. This means that the selected image may be close to other impression words. Therefore, it may cause a sense of incongruity by displaying an image with a different concept. In the future, it is necessary to increase the number of images to ensure that there is always an image that matches the concept.

**TABLE VI:** Concept of image

| Image | Image label | Most frequent color (R,G,B) | Concept (R,G,B) |
|---|---|---|---|
|  | pink room | (182,78,100) | (192,0,112) |
|  | five-storied pagoda | (159,44,19) | (104,0,31) |
|  | strawberry | (106,120,81) | (107,107,98) |
|  | fireworks | (7,9,7) | (0,60,60) |
|  | baby blue eyes | (73,166,227) | (91,189,206) |

(3) It would be insufficient the number of impression words and colors in the shared concept space. In this study, we have set 27 categories of impression words based on CIS. By increasing the number of impression words and colors, it would be expected higher accurate classification in shared concept space. Furthermore, CIS links the impression words of emotion and sensibility to colors based on a standard human perception of color. Therefore, since adjectives are set as impression words rather than nouns or verbs, there may be issues in similarity word

**TABLE VII:** Keyword and displayed image:shinning star

| Displayed image | Keyword | Impression word |
|---|---|---|
|  | wind | mild |
|  | dream | cute |
|  | world | nature |

**TABLE VIII:** Keyword and displayed image:Diamond Dust-tribute to Frozen-

| Displayed image | Keyword | Impression word |
|---|---|---|
|  | scenery | romantic |
|  | proof | safety |
|  | beside | stay |

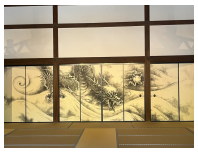
calculations using Word2Vec.

Additionally, as mentioned in the comments, it may enhance our proposed system by displaying lyrics on the images.

## IV. Conclusion

In this paper, we proposed a method to automatically generate slideshow synchronized with song by using concepts for lyrics and images, and utilizing their proximity. The experimental results demonstrated that by extracting and matching concepts between lyrics and images, we confirmed that slideshows synchronized with music are generated, which displays images that matched the lyrics at the time of their occurrence. However, there are several areas for improvement, such as refining the shared concept space and the method of image concept extraction, to achieve better synchronization with music and higher slideshow quality. In addition, it is also important future task to find methods for quantitative evaluation.

**TABLE IX:** Keyword and displayed image:Success!

| Displayed image | Keyword | Impression word |
|---|---|---|
|  | tomorrow | gentle |
|  | hand | strong |
|  | match | clear |

## References


[1] Hiromi Ishizaki et al. “Music Slideshow Generation Based on Web Image Retrieval with Queries Constructed from Lyrics (in Japanese)”. In: *IPSJ Journal* 54.4 (2013), pp. 1263–1274.

[2] Shigenobu Kobayashi. *Color Image Scale*. Distributed in the U.S. by Kodansha America, 1992.

[3] Nozomi Midorikawa, Toru kano, and Yuriko Takeshima. “Music Visualization Based on Lyrics and Acoustic Features (in Japanese)”. In: *ITE Technical Report* 41.12 (2017), pp. 271–272.

[4] Ken Murata et al. “A method for recommending songs based on time series similarity based on the story development of lyrics (in Japanese)”. In: *Proceedings of 84th National Convention of IPSJ*. 2022, pp. 525–526.

[5] Ryutaro Niho and Yasuyuki Saito. “A study of automatic recommendation system of music matching image impesslon”. In: *ITE Technical Report Vol.37, No.7*. 2013, pp. 23–26.

[6] Yasunori Ohishi et al. “Conceptbeam: Concept driven target speech extraction”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 4252–4260.