Research Article

Ningjie Liao*

Research on intelligent interactive music information based on visualization technology

https://doi.org/10.1515/jisys-2022-0016
received November 02, 2021; accepted December 29, 2021

Abstract: Combining images with music is a music visualization to deepen the knowledge and understanding of music information. This study briefly introduced the concept of music visualization and used a convolutional neural network and long short-term memory to pair music and images for music visualization. Then, an emotion classification loss function was added to the loss function to make full use of the emotional information in music and images. Finally, simulation experiments were performed. The results showed that the improved deep learning-based music visualization algorithm had the highest matching accuracy when the weight of the emotion classification loss function was 0.2; compared with the traditional keyword matching method and the nonimproved deep learning music visualization algorithm, the improved algorithm matched more suitable images.

Keywords: music visualization, convolutional neural network, long short-term memory, synesthesia

1 Introduction

Music is an acoustic way of expressing emotional thoughts, and it is also an art form. Music uses changes in the rhythm and pitch of sounds to convey information. When people receive such information, they can not only appreciate the rhythm and melody, but also feel the change of emotion [1]. With the improvement of living standards, people's demands for spiritual culture are also getting greater. People are no longer satisfied with listening to music, but want to "see" the emotional changes in music while listening. Music visualization can connect people's auditory and visual senses, so that people can feel the information contained in music more intuitively [2]. In addition to enhancing the appreciation of music, music visualization can also show the sound characteristics that are not intuitive enough and need to be subjectively perceived by people in a more accurate and straightforward way, and visually observing the sound characteristics of music can assist in music teaching or composition [3]. Plewa and Kostek [4] proposed a graphical representation method of song emotions based on self-organizing mapping and created a map in which music excerpts with similar moods were organized next to each other on the two-dimensional display. Li and Li [5] have constructed a music visualization model based on graphic images and mathematical statistics by combining mathematical, statistical methods such as K-mean clustering and fusion decision trees based on music graphic images to address the current shortcomings in the field of music visualization. The actual case analysis and performance test results showed the superiority of the music visualization method based on graphic images and mathematical statistics. Lopez-Rincon and Starostenko [6] proposed a method to normalize data in musical instrument digital interface files by 12-dimensional

* **Corresponding author: Ningjie Liao,** College of Music and Dance, Hunan University of Science and Engineering, No. 130, Yangzitang Road, Chaoyang Office, Lingling District, Yongzhou, Hunan 425199, China, e-mail: liaobutan5kj@163.com

vector descriptors extracted from tonality and a novel technique for dimensionality reduction and visualization of extracted music data by three-dimensional projections. They found through experiments that the method retained 90% of the original data in the dimensionality reduction projection. The literature mentioned above show that feature extraction of music data is important for music visualization. In this article, convolutional neural network (CNN) and long short-term memory (LSTM) were used to extract features from images and music more accurately, and the images were combined with music. This study provides an effective reference for music visualization. This article studied the combination of music and images to make audience understand the information contained in music more comprehensively. The novelties of this study are that music was paired with images using CNN and LSTM, and the sentiment classification loss function was added to the traditional loss function to make the algorithm consider the sentiment contained in music and images during training and achieve the proper matching of music and images.

2 Music visualization

Music visualization, in a narrow sense, is the visualization of the sound characteristics of music, such as the time-frequency diagram of sound, but in a broader sense, it is the interpretation of the information contained in music through pictures or videos, providing an intuitive visual presentation to users. This study focuses on the visual representation of the emotional information contained in music in music visualization. Broadly speaking, music visualization uses images to visually interpret the content of music, and its principle lies in synesthesia [7]. The literary rhetoric of synesthesia refers to the fact that a stimulus to one sense evokes the perception of that sense and the perception of another sense. Music visualization mentioned in this article is audio–visual synesthesia, which refers to the visual association caused by auditory perception or auditory association caused by visual perception. Among the various human senses, vision is arguably the most important channel for receiving information, and when visual and auditory senses produce a synesthesia effect, auditory perception is enhanced so that more accurate judgments can be made when receiving external information [8].

Music visualization technology is also a classification of interaction technology, where “interaction” is defined as an action between plural objects that affect each other. Interaction can be divided into human–human interaction, human–computer interaction, and computer–computer interaction. Music visualization technology is human–computer interaction. The user inputs music into the computer, and the computer processes the music through the corresponding algorithm and outputs it to the user after stitching it together with the corresponding images, which affects the user from both auditory and visual aspects [9].

3 Deep learning-based music visualization

3.1 Music and image matching based on a deep learning algorithm

Traditionally, music is visualized by means of a weak emotional label that corresponds to some of the characteristics of the music. However, on the one hand, the emotion weak tags of the images have the wrong emotion descriptions, and on the other hand, the emotional information of music often needs to be fully expressed through the whole, which means that the emotion expressed by music needs to be contextualized [10]. Due to the above two reasons, traditional music-image matching methods are not effective for music visualization. Therefore, this study uses a deep learning algorithm to pair music signals and images to visualize the sentiment information in music signals. For the candidate images, CNN is used for feature extraction; for the music signal, as its emotional information features need to be related to the context, that is, the order of information affects the feature expression, a variant of recurrent neural network (RNN) – LSTM is used to extract the music signal before feature-based pairing. Figure 1 shows the basic flow of the above deep learning-based music emotion visualization.

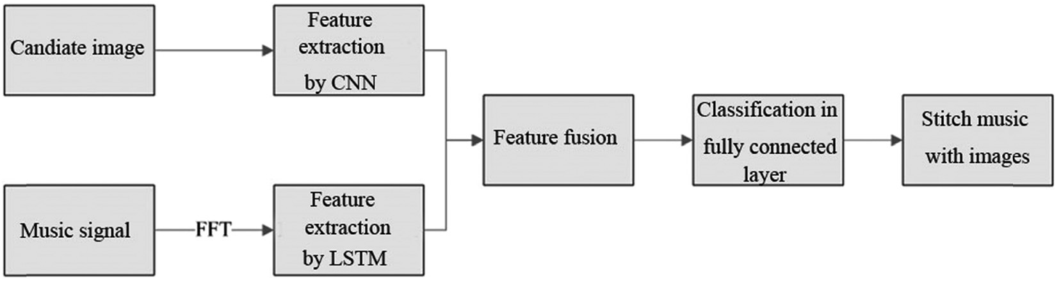

Candiate image
Feature extraction by CNN
Music signal
FFT
Feature extraction by LSTM
Feature fusion
Classification in fully connected layer
Stitch music with images

Figure 1: Deep learning-based music visualization process.

1. After preprocessing the candidate images, feature extraction is performed using CNN [11]. The convolution and pooling layers are interleaved, that is, one convolution layer and one pooling layer, or multiple convolutional layers and one pooling layer. The image features after multiple convolutional kernel pooling operations are used for subsequent matching classification.
2. The music signal is preprocessed, and then the fast Fourier transform [12] is performed on the signal to obtain the spectrogram. The emotional semantic features of the spectrogram are extracted using LSTM.
3. The candidate image features extracted by CNN and the sentimental features of the music signal extracted by LSTM are stitched together to obtain the fused features.
4. The fused features of the candidate image and music signal are input into the fully connected layer for classification to judge whether the music signal matches the candidate image. If they do not match, the candidate image is replaced, and feature extraction and judgment are performed again; if they match, the music and image are stitched together according to the timeline of the music [13].

The above steps are the basic steps for the practical application of the music visualization algorithm after training. Samples with corresponding pairing labels are input into the algorithm during training, and the same steps 1–3 are followed for feature extraction and feature fusion of the samples. Then, the fused sample features are classified to determine whether they match. The classification results are compared with the actual classification labels of the training samples. For the error calculation of the classification results, cross-entropy is often used to calculate the loss function, and then the calculated loss function is used to reversely adjust the structural parameters.

To further improve the accuracy of matching and enhance the emotional embodiment of matched images to deepen listeners' understanding of musical emotions, this study improves the loss function by introducing emotional information labels to supervise the training. The formula of the improved loss function is as follows:

$$\begin{cases} \text{Loss} = \text{Loss}_1 + \lambda\ \text{Loss}_2, \\ \text{Loss}_1 = -\sum c_i \log(p_i), \\ \text{Loss}_2 = -\sum c_i' \log(p_i'), \end{cases} \tag{1}$$

where Loss, Loss_1, and Loss_2 are the overall matching loss, music and image matching loss, and music and image sentiment classification loss, respectively, λ is the weight of sentiment classification loss, c_i and c_i' are the true labels of music and image matching and sentiment matching, respectively, and p_i and p_i' are the music and image matching probability and sentiment matching probability, respectively.

3.2 Synthesis of music and images

In the previous subsection, the matching algorithm of music segments and pictures was introduced and improved. After the matching of musical phrases and pictures is completed, it is necessary to synthesize the musical phrases and pictures into music videos to visualize the music, and the basic flow is shown in Figure 2.

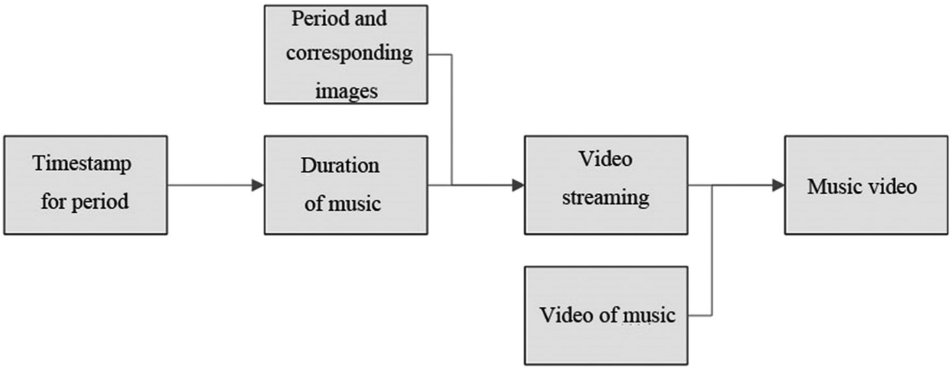


Figure 2: The synthesis process of music and images.

1. The period in the music file is marked with a timestamp label.
2. The duration of the music is calculated based on the timestamp label of the period, and the duration that the matched image corresponding to the period can last is calculated.
3. The images corresponding to the period are clipped into a video stream in the order of the timestamp of the period. The duration of every image in the video stream is calculated by step 2.
4. The periods corresponding to the images are synchronized with the video stream to obtain a music video to realize music visualization.

4 Experimental analysis

4.1 Experimental environment

In this study, simulation experiments were conducted in a laboratory server using MATLAB software (The MathWorks, Inc., Natick, Massachusetts, USA) [14] with the following relevant configurations: Windows 7 operating system, Core i7 processor, and 32 GB memory.

4.2 Experimental data

The deep learning-based music emotion information visualization algorithm can realize intelligent human–computer interaction between music, images, and people. After the user input music into the algorithm, it matched with the candidate images, and the matched music and images were stitched together and fed back to the user, so that the user could more deeply experience the emotion expressed in the music. Before using this visualization algorithm, the algorithm was trained by training samples in order to match the music with images accurately.

A crawler written in Python crawled 1,500 English songs with lyrics from music platforms. The periods of these songs were divided according to the numbered musical notation. Then, images were retrieved by the Baidu search engine with the aid of the lyrics of the periods. The top 20 images retrieved according to the lyrics of each period were taken as the candidate images. Then, period and candidate image pairings were initially screened using voting and finally constructed by manual scoring and voting, and the pairs were marked with sentiment labels. Finally, constructed period–image pairs with sentiment labels were used as positive samples, and the period–image pairs composed of period and other candidate images without sentiment labels were used as negative samples. The ratio of positive samples to negative samples was 1:3. The basic data formats in the samples were jpg format for the image and mp3 format for the period. The names of the images and periods paired in the positive samples were the same. Every pair had a json file in which the sentimental type and song name of the pair were recorded, and the name of this file was the same as the name of the pair.

4.3 Experimental setup

In the deep learning-based algorithm for visualizing music emotional information, the parameters of the CNN used to extract image features are as follows. There were 13 convolutional layers and 5 pooling layers. The structural distribution of the two kinds of layers was two convolutional layers – one pooling layer – two convolutional layers – one pooling layer – three convolutional layers – one pooling layer – three convolutional layers – one pooling layer – three convolutional layers – one pooling layer. The convolutional layers had 64 convolutional kernels in a size of 5 × 5. The Relu function was used as the activation function. The pooling layer adopted max-pooling. The size of the pooling frame was 2 × 2. The sliding step length was 2. As to LSTM used for extracting music features, a four-layer bidirectional RNN was used as the hidden layer, which contained long and short-term memory units such as forgetting gates, recurrent gates, and output gates as described earlier.

In the improved loss function used for supervising algorithm training, the weight λ of the sentiment classification loss would affect the matching accuracy of the whole algorithm. The value of λ was set as 0, 0.2, 0.4, and 0.8, respectively, in this study. The accuracy of the trained algorithm under the four weights was tested.

To further verify the matching accuracy of the algorithm, it was compared with the traditional music matching method. The traditional music matching method matched images and music based on the weak tags of the images and the keywords contained in the lyrics of the period. During the comparison, the sentiment classification loss weight λ of the algorithm was set as 0 and 0.2, respectively; the former represented the visualization algorithm before the improvement of the loss function, and the latter represented the visualization algorithm after the improvement of the loss function. The reason for selecting 0.2 was that 0.2 was the optimal weight obtained in the previous weight test.

4.4 Evaluation indicators

Accuracy under $R@K$ [15] was used as the evaluation indicator of the music visualization algorithm, and the candidate images matched with the periods were ranked according to the matching probability calculated by the algorithm from the largest to the smallest. The matching was considered as successful when the first K images contained the correct images. The value of $R@K$ meant the proportion of the successfully matched periods under the K value. The value of K was set as 1, 5, and 10. The reason for using accuracy under $R@K$ as an evaluation criterion is that the information contained in the music and the information conveyed by the images only partially overlapped unless a specific figure was drawn for the music. The difference was the degree of overlapping. Therefore, there was more than one matching result when matching a music with the image bank. In addition, as the sentimental information of music images was strongly subjective and difficult to be quantified, different values of K meant the visualization algorithm could give K candidate images that were closest to the period for the user to select, improving human–machine interaction.

4.5 Experimental results

In this study, the loss function used for supervising training in the deep learning-based music visualization algorithm was improved to make full use of emotional information in the periods and images. Sentiment classification loss was added to the loss function, and the proportion of the sentiment classification loss was adjusted by the weight. The accuracy of the improved sentiment visualization algorithm with different weight proportions is shown in Figure 3.

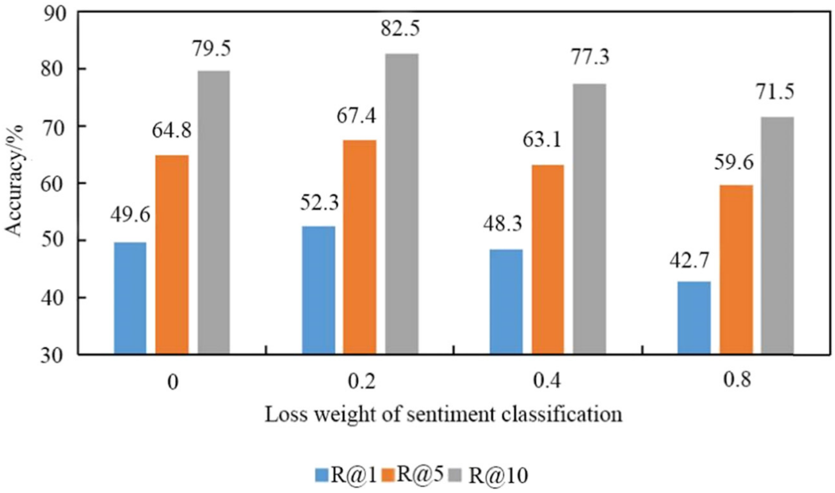
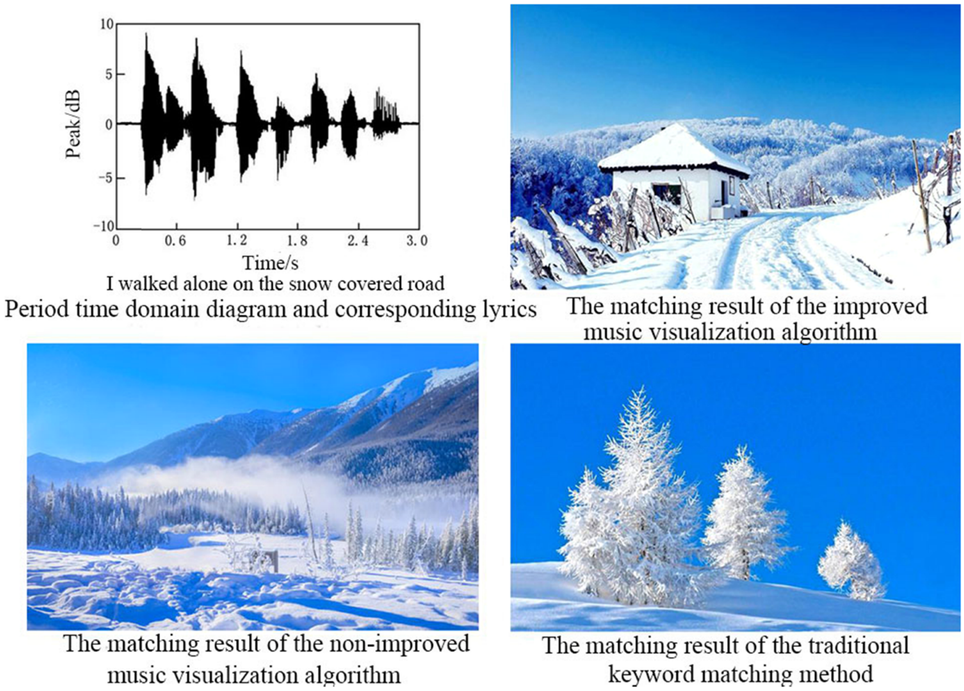


Figure 3: Accuracy of the improved music visualization algorithm under different loss weights of sentiment classification.

Figure 3 shows that the matching accuracy of the visualization algorithm increased with the increase of the value of K in $R@K$when the loss weight of sentiment classification was the same. The reason for the above result is that the user's feeling on the music and image was subjective; the smaller the value of K was, the smaller the available range for matching was, the more difficult the matching was. In addition, comparing the matching accuracy of the visualization algorithm with different loss weights of sentiment classification under the same $R@K$. Figure 3 showed that the visualization algorithm had the highest accuracy when the weight was 0.2. The reason for the above result is that the sentiment information in the periods and images could assist in matching the periods with the images more accurately, thus making the music visualization more accurate, but the sentiment labeling of the sample set used in the algorithm training was determined by using the voting score, which had errors. The larger the loss weight of sentiment classification was, the larger the influence on the matching results was.

Limited by space, this article only showed partial matching results of the three music visualization algorithms under $R@1$, as shown in Figure 4. Figure 4 provides the time-domain diagram of the period, but the sentiment information cannot be directly felt from the time-domain diagram alone due to the text format; therefore, the lyric of the period was used to assist in the illustration. It was seen from the lyric shown in Figure 4 that the period described a person walking alone on a snowy road, and the peak in the time-domain diagram of the period was gradually decreasing to reflect the emotion of loneliness and desolation. The image given by the traditional keyword and music matching method only showed a few



Period time domain diagram and corresponding lyrics

The matching result of the improved music visualization algorithm

The matching result of the non-improved music visualization algorithm

The matching result of the traditional keyword matching method

Figure 4: Partial matching results of three music visualization algorithms.

snow-covered trees on the snow, and there was no significant intersection except for the overlap between the element of “snow” and the lyrics of the period; the image given by the nonimproved music visualization algorithm had a larger area of snow and trees, and the various marks on the snow also reflected the element of “road”; the improved music visualization algorithm gave an image of a snowy road, with snow on both sides of the road, visible ruts in the snow, dead branches on both sides of the road, and small houses with snow on one side, which not only reflected the “road” but also embodied a sense of loneliness.

To further verify the matching performance of the improved deep learning-based music visualization algorithm for periods and images, it was compared with the nonimproved deep learning-based music visualization and the traditional keyword matching method, in which the improved music visualization algorithm used a loss weight of 0.2 for sentiment classification. The comparison results are shown in Figure 5.

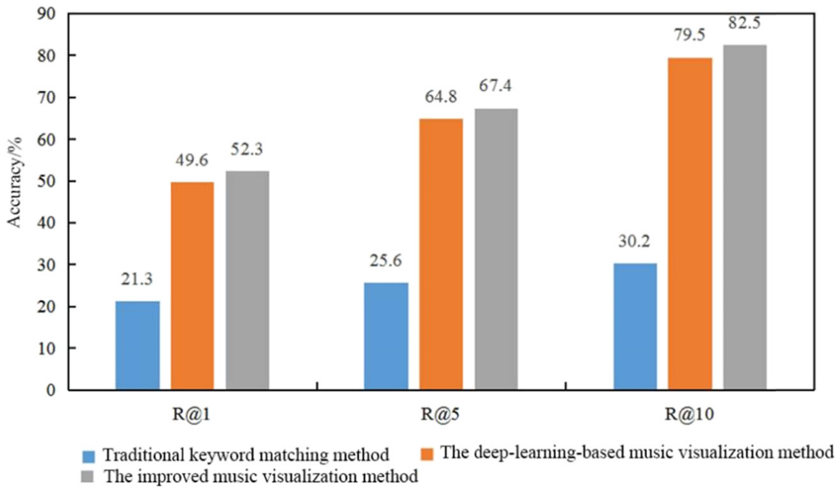


Figure 5: Matching accuracy of three music visualization algorithms.

Figure 5 shows that the improved deep learning-based music visualization algorithm had the highest matching accuracy under the same *R@K*, the nonimproved deep learning-based music visualization algorithm had the second highest matching accuracy, and the traditional keyword matching method had the lowest. In addition, with the increase of the value of *K* in *R@K*, the matching accuracy increased no matter what kind of music visualization algorithm was used.

5 Conclusion

This study introduced the concept of music visualization, paired music and images with CNN and LSTM for music visualization, and added the sentiment classification loss function to the loss function to make full use of the emotional information in music and images, and finally conducted simulation experiments. The results are as follows: (1) as the value of *K* in *R@K* increased, the candidate range of matching increased, the matching difficulty decreased, and the accuracy of the improved deep learning-based music visualization algorithm increased accordingly. (2) The improved deep learning-based music visualization algorithm had the highest accuracy when the loss weight of sentiment classification was 0.2. (3) The improved deep learning-based music visualization algorithm matched images more closely to music than the traditional keyword matching method and the nonimproved deep learning music visualization algorithm. (4) The improved deep learning-based music visualization algorithm had the highest matching accuracy under the same *R@K*; the matching accuracy increased as the value of *K* in *R@K* increased no matter what kind of music visualization algorithm was used.

The future research direction is to expand the range of training samples to improve the accuracy of music and image matching and offer an effective reference for music visualization technology.

Conflict of interest: Author states no conflict of interest.

References

[1] Chen X, Ashoor H, Musich R, Wang J, Zhang M, Zhang C, et al. Epihet for intra-tumoral epigenetic heterogeneity analysis and visualization. Sci Rep. 2021;11:376.

[2] Gupta S, Das SK, Jana D, Pal DK. Distraction during cystoscopy to reduce pain and increase satisfaction: Randomized control study between real-time visualization versus listening to music versus combined music and real-time visualization. Urol Ann. 2019;11:449.

[3] Gaulon C, De Rec C, Combriat T, Marmottant P, Elias F. Sound and vision: visualization of music with a soap film, and the physics behind it. Eur J Phys. 2017;38:045804.

[4] Plewa M, Kostek B. Music mood visualization using self-organizing maps. Arch Acoust. 2015;40:513–25.

[5] Li W, Li J. Research on music visualization based on graphic images and mathematical statistics. IEEE Access. 2020;8:1.

[6] Lopez-Rincon O, Starostenko O. Music visualization based on spherical projection with adjustable metrics. IEEE Access. 2019;7:1.

[7] Xu S. Research on the visualization of music stage performance based on the context of computer digital media. J Phys Conf Ser. 2021;1915:022027.

[8] Chaturvedi V, Kaur AB, Varshney V, Garg A, Chhabra GS, Kumar M. Music mood and human emotion recognition based on physiological signals: a systematic review. Multimedia Syst. 2021;28:1–24.

[9] Venkatarangam S. Participants' experiences of a receptive music therapy intervention that incorporates raga: an interpretative phenomenological analysis. Art Psychother. 2021;73:101762.

[10] Shen J, Wang R, Shen HW. Visual exploration of latent space for traditional Chinese music – ScienceDirect. Visual Inform. 2020;4:99–108.

[11] Dhiraj R, Biswas, Ghattamaraju N. An effective analysis of deep learning based approaches for audio based feature extraction and its visualization. Multimed Tools Appl. 2019;78:23949–72.

[12] Song H, Yang Y. Super-resolution visualization of subwavelength defects via deep learning-enhanced ultrasonic beamforming: a proof-of-principle study. NDT E Int. 2020;116:102344.

[13] Yang C, Li Y, Liu C, Yuan X. Deep learning-based viewpoint recommendation in volume visualization. J Visual. 2019;22:991–1003.

[14] Mauch L, Wang C, Yang B. Subset selection for visualization of relevant image fractions for deep learning based semantic image segmentation. J Franklin I. 2018;355:1931–44.

[15] Castillo JR, Flores MJ. Web-based music genre classification for timeline song visualization and analysis. IEEE Access. 2021;9:18801–16.