Design and Implementation of a Cross-Media Image to Sound Algorithm and Objective Evaluation Algorithm

Zihao Liu and Botao Wang
Faculty of Information Technology, Beijing University of Technology, Beijing, China
Email: 1687240670@qq.com; wangbt@bjut.edu.cn

Abstract. With the joining of digital media and Internet, the application of the conversion of cross-media information becomes wider and wider, which makes the integration of multi-media information possible. The transformation of image information to sound information is also in the stage of exploration. It can be applied to all aspects of the industry. Most importantly, it can be used to guide the blind. In this paper, an improved image-to-sound algorithm based on original method is proposed to judge whether there are obstacles in front of the road and test the sound segment through human ears. Among different mapping methods, the highest accuracy can reach 84.2%, which is much higher than the original method. At the same time, an objective evaluation algorithm is designed based on the neural network, which can replace the human ear to judge a large number of image information, and the matching degree of this model can reach 68%.

Keywords. Cross-media information; guide the blind; image-to-sound; objective evaluation.

1. Introduction
Media plays a vital role in human communication, through which we can get the information we want. In the past, the types of media were very limited. Later, with the development of media and related technologies, the convergence between media has become stronger and stronger, which also brings a variety of possibilities of media [1]. The expression form of media is also developing towards multi-medium and multi-dimensional [2]. The convergence of text, image, video and sound media is becoming more and more common, making cross-media information a qualitative leap and a great breakthrough for every walk of life. For example, in education, teachers can use multimedia to deal with software more conveniently [3]. Similarly, in terms of music, graphic design is made according to the data of different timbres and tones of each song to form graffiti paintings [4]. In addition, the problem of singing is analyzed according to the spectrum diagram of sound, so as to achieve the purpose of vocal music teaching [5].

Cross-media information can also be applied to the visually impaired. Some of them can’t see the world as it is from birth, while others become blind after birth. They can’t see all kinds of changes in the world and can only hear the world through their ears. With the improvement of medical treatment in recent years, some people who were blinded by the disease regained their sight. However, some people's vision is still in the dark, so helping these visually impaired people to better adapt to life has become a relatively important topic. In this paper, an improved image to sound method is proposed. After the image of the road ahead is transformed into sound, and the human ear is used to judge the presence of obstacles on the road ahead, the accuracy rate is significantly improved. At the same time, this paper also proposes an objective evaluation algorithm, which can replace the human ear to judge the large amount of image data, and the accuracy of the judgment is similar.
2. The Original Image to Sound Method

The general algorithm principle and process are shown in figure 1 [6]. The basic principle of this original image-to-sound method is to directly reflect the information of gray image pixels into the sound. The basic process is image of gray first, convert each column of pixels from image to sound, and then assemble all the columns according to the sequence of time. In the transformation of each column, the gray value of each pixel is reflected to the amplitude of sound sine wave, and the position of the pixel in the image row is reflected to the frequency of sine wave, and then the sine wave corresponding to a column of pixels is superimposed. The frequency formula of the pixel at each position is as follows [7]. The pixel frequency will vary depending on the location.

\[ f = \frac{\text{Pixel frequency}}{\text{Sampling frequency}} \]  

(1)

3. Improved Image to Sound Method

This is an attempt of a new method, which eliminates the steps of image of gray and is inspired by the method of image to music [8]. This method firstly obtains the RGB value of each pixel of a color picture, and then obtains the HSV value of each pixel through the conversion formula of RGB to H (hue), S (saturation) and V (lightness). The transformation formula is as follows:

\[ \max = \max (R, G, B) \]  

(2)

\[ \min = \min (R, G, B) \]  

(3)

\[ V = \max (R, G, B) \]  

(4)

\[ S = \frac{\max - \min}{\max - \min} \]  

(5)

\[ H = \begin{cases} 
\frac{G - B}{\max - \min} \times 60 \ (R = \max) \\
\frac{B - R}{\max - \min} \times 60 + 120 \ (G = \max) \\
\frac{R - G}{\max - \min} \times 60 + 240 \ (B = \max) \\
H + 360 \ (H < 0) 
\end{cases} \]  

(6)

After the three values of HSV are obtained, the initial phase is added on the basis of the original method, so that the three values can be corresponding to the three variables of the sine wave. First, the value of H corresponds to the amplitude of the sine wave, the value of S corresponds to the angular
frequency of the sine wave, and the value of V corresponds to the initial phase of the sine wave. After obtaining the sine wave corresponding to each pixel, the sine waves are accumulated in columns, and then presented in chronological order from left to right. By using the new method, the correct rate of judging whether the road ahead has obstacle or not can reflect the effect of this method. Then all the mapping cases are enumerated to get the accuracy of the judgment and find a better correspondence.

4. Objective Evaluation Algorithm

When the picture data is small, it is more appropriate to judge by the human ear. However, when the picture data quantity becomes large, if the human ear listens in turn, the efficiency will be very slow. Moreover, listening for a long time will affect the result of judging the picture. Therefore, an objective evaluation algorithm is designed according to the neural network. By establishing a model, the judgment of the model on the sound is as similar as that of the human ear. Thus, more sound information can be judged by this model, and the result of whether there are obstacles ahead can be further obtained. This objective evaluation will also provide a lot of convenience for the following research.

4.1. Sample Data Acquisition

The sound sample data in this paper are all the sounds mapped above and then processed to obtain sound files suitable for this objective evaluation algorithm. In addition, due to the need to use the neural network model, the data set needs to be increased, so the method of data enhancement is used to expand the image data [9], and then the image-to-sound algorithm is used to convert it into sound files. At the same time, it can prevent the phenomenon of overfitting. Three methods of original image processing are selected here, which are contrast increase, brightness increase and flip horizontal.

In this experiment, all the sound samples were labeled, a total of 1652 short audio segments (each segment of audio is less than 4s) were divided into two categories, as shown in table 1.

| Number | Type       | Number of voices |
|--------|------------|------------------|
| 0      | barrier    | 736              |
| 1      | barrier-free| 916              |

4.2. Selection of Characteristic

4.2.1. MFCC. MFCC is proposed based on the audio characteristic of human ears [10], and it has a nonlinear corresponding relationship with HZ frequency. Extracting MFCC from sound signals requires pre-emphasis, Sub-frame, plus windows, FFT and other operations. Because of its characteristics, MFCC is widely used in speech signal recognition and analysis.

4.2.2. Tonnetz. This is a kind of musical grid diagram used to represent tonal space [11]. In this way, the distance and relationship between tones are represented in space, and the relationship between tones in sound is analyzed in a visual way.

4.2.3. Chrominance Frequency. The frequency spectrum of this sound is projected into 12 Spaces [12], and the octave space of music is divided into 12 different musical scales with chromatic units in accordance with twelve-tone equal temperament. The chromaticity feature is the distribution of these scales in the music. Therefore, this feature can also well reflect the pitch size and other characteristic information of the voice.

4.2.4. Mel-spectrum. When it is difficult to see the characteristics of sound in time domain, FFT will generally be used to analyse the characteristics of the spectrum. However, there is lack of time information in the spectrum. STFT method can take into account the time and frequency at the same
time in order to obtain more complete sound information. However, the sonograph is usually very large, so in order to get the sound features of appropriate size, the Mel- spectrum is generally obtained through the transformation of Mel Filter Banks [13].

4.3. Code
One-hot coding is used here [14], which is a common method in classification problems, and N-bit state register is used to encode N states. Different categories are represented by different binary vectors. After mapping different variables into integer values, the following characteristics will be marked as 1 if they fit into the category, and the rest values will be 0. All the data will be stored in a matrix after coding, which provides convenience for the calculation of loss function and accuracy rate later.

4.4. Design of Neural Network
After processing the above characteristic information, it is used as the input layer of the neural network, and then two layers of fully connected layer are added as the hidden layer. After each full connection layer, two layers of dropout layer are added to prevent data over-fitting. The activation function of the output layer is Softmax for classification [15]. The rest of the activation functions are set to RELU. The network loss function uses the cross-entropy loss function, and the optimizer chooses Adam initially. The batch_size is set to 64 and the iteration period is set to 100. Then you can train the data.

5. Experiments and Result Analysis
5.1. Improved Image to Sound Algorithm
There were 40 test pictures, 20 of which were with obstacles in front and 20 of which were without obstacles in front. The following different mapping methods were used to obtain 40 audio segments, a total of 240 audio segments. For each mapping method, 21 data are first listened to by human ears as learning data, and then the remaining data are listened to as testing data. The judgment accuracy obtained by different mapping methods is shown in the following table 2.

| The mapping relation of H,S,V | Determine the correct number | Number of judgment errors | Accuracy (%) |
|-------------------------------|------------------------------|---------------------------|--------------|
| A,ω,φ                        | 13                           | 6                         | 68.4         |
| A,φ,ω                        | 12                           | 7                         | 63.2         |
| ω,A,φ                        | 15                           | 4                         | 78.9         |
| ω,φ,A                        | 12                           | 7                         | 63.2         |
| φ,ω,A                         | 11                           | 8                         | 57.9         |
| φ,A,ω                         | 16                           | 3                         | 84.2         |

In this experiment, the original gray scale degradation method of Image to Sound was improved. The RGB value of each pixel of the image was directly converted into HSV value, and the HSV value was mapped to the amplitude, angular frequency and initial phase of the sine wave. Finally, the waveform was obtained through column superposition, and then the waveforms were arranged in time order to obtain the final total waveform. Compared with the original Image to Sound method, this method adds a variable named initial phase on the basis of the original one, so that the sine wave obtained by each pixel has more characteristics of the original image pixel. In addition, H,S,V color space is used in this experiment, which reduces the huge calculation amount of RGB three-channel. Besides simplifying, the color distribution of the original image can be perfectly reflected, which is convenient for judgment. Moreover, in the judgment of obstacles in the disordered picture test, the number of times of listening to each segment of audio was significantly reduced, and the
characteristics of the judgment were much clearer than before. The original method needed 3 times on average to make the judgment. After the improvement of the method, the accuracy of some mapping methods is higher than that of the original method, which indicates that the best effect can be obtained when the image saturation (S) is mapped to the amplitude of sine wave. The rest of the mappings are more or less accurate than before, even if they are less accurate.

5.2. Objective Evaluation Algorithm
The data source expands the data by using image enhancement method on the basis of taking photos by oneself, and converts the data into audio by using the image-to-sound method mentioned above. The training set consists of 1200 audio clips, the verification set consists of 400 audio clips, and the rest are used as the test set. The training results are as follows.

![Figure 2. Training error curve.](image)

![Figure 3. Training and validation curves.](image)

It can be seen from figure 2 that the loss function of the training curve continues to decrease and reaches a low value, which meets the standard of model training. It can be seen from figure 3 that the matching degree between this evaluation system model and human ear can basically reach 68%. This objective evaluation is compared with the single feature method that has been tried before, the results of this objective evaluation are shown in the following table 3. The accuracy has also been improved, indicating that although the calculation amount of multiple characteristic parameters has been increased, the acquisition of audio information is more detailed, and multiple criteria can be used to judge at the same time to better distinguish the two groups of whether there is obstacle or not.

| Method                  | Accuracy (%) |
|-------------------------|--------------|
| Kmeans                  | 60           |
| Single feature MFCC     | 55           |
| Objective evaluation of this paper | 68           |

6. Conclusion
Aiming at the problem of judging whether there are obstacles ahead, an improved image to sound method is studied and proposed. According to the different mapping results, multiple results are obtained. The accuracy rate of some results is as high as 84.2%, which is higher than the original image to sound method. At the same time, an objective evaluation algorithm is designed, which can replace the human ear to complete the judgment of large data volume. The neural network model obtained four kinds of sound features as input, and finally the matching rate between the model and human ear could reach 68%. The accuracy rate is higher than the previous single feature evaluation. In
the future, in the aspect of image-to-sound method, we will continue to explore better improved methods to reduce subjective inference and prevent everyone from having different understandings of a segment of audio. Secondly, the model of objective evaluation also has a lot of room for improvement. Later, some model parameters, such as learning rate and loss function, will be modified to try to improve the model matching degree with higher accuracy.

References
[1] Qu G T 2019 Research on Media Transformation of Chinese Character Graphics under Digital Media Art (Nanjing University of the Arts).
[2] Sun W 2007 Dynamic Text Design in Digital Media (Nanjing University of the Arts).
[3] Xu Y G and Zhao L 2007 Application research of limit download technology for multimedia material resources China Medical Education Technology (04) 83-85.
[4] Pang L 2013 Approaches and methods to transform from music to vision Journal of Nanjing Arts Institute (Fine Arts & Design) (01) 149-155.
[5] Wu D 2018 Vocal music carrier is transformed from subjective feeling to objective image--using spectrograph to analyze human voice for vocal music teaching Northern Music 38 (9) 142-143.
[6] Peng Y Q, Gao J J, Liang C J and Li T J 2013 Research on audiovisual information fusion method based on image audible Computer Applications and Software 30 (11) 76-79.
[7] Tian Y N, Lei H W and Wang X 2010 Automatic blind guide system based on SVM image segmentation method Journal of Northeastern University (Natural Science) 31 (12) 1681-1684.
[8] Wang B T, Liu Y H and Li X M 2017 A summary of blind guiding methods converting images to sounds 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).
[9] Geng P Z, Tang Y Q, Fan H X, Zhang S R and Zhu X T 2021 Research on the effect of data enhancement on deep forgery detection model Computer Engineering and Applications 1-10.
[10] Chen B and Yu Y Y 2019 Study on urban sound classification model based on deep neural network Journal of Zhejiang University of Technology 47 (02) 199-203.
[11] Harte C and Sandler M 2006 Detecting harmonic change in musical audio Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia pp 21-26.
[12] Zhang H Y 2020 Research on Carrier-less Information Hiding Technology Based on Computer Aided Composition (Beijing University of Posts and Telecommunications).
[13] Liu L J 2015 Research on Neural Network Modeling Method for Sound Conversion (University of Science and Technology of China).
[14] Gu B and Sung Y 2021 Enhanced reinforcement learning method combining one-hot encoding-based vectors for CNN-based alternative high-level decisions Applied Sciences 11 (3) 1291.
[15] Peng X B, Huang H N, Yang H Y, Liu J H and Huang Y 2021 Gait recognition method based on walking feature vector image Journal of Shenzhen University (Science and Engineering) 1-8.