Learning cross-modal visual-tactile representation using ensembled generative adversarial networks

Xinwu Li¹, Huaping Liu², Junfeng Zhou², FuChun Sun²

¹School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068, People’s Republic of China
²Department of Computer Science and Technology, Tsinghua University, Beijing 100084, People’s Republic of China

E-mail: hpliu@tsinghua.edu.cn

Abstract: In this study, the authors study a deep learning model that can convert vision into tactile information, so that different texture images can be fed back to the tactile signal close to the real tactile sensation after training and learning. This study focuses on the classification of different image visual information and its corresponding tactile feedback output mode. A training model of ensembled generative adversarial networks is proposed, which has the characteristics of simple training and stable efficiency of the result. At the same time, compared with the previous methods of judging the tactile output, in addition to subjective human perception, this study also provides an objective and quantitative evaluation system to verify the performance of the model. The experimental results show that the learning model can transform the visual information of the image into the tactile information, which is close to the real tactile sensation, and also verify the scientificity of the tactile evaluation method.

1 Introduction

With the rapid development of technology and success of artificial intelligence, object material recognition has been extensively used in many industrial fields, such as e-commerce, machinery manufacturing, and intelligent robots [1]. At this stage, the distinguishing method of traditional object materials still stays in the visual difference discrimination. However, due to the influence of environmental factors such as illumination and temperature, and the limitations of the imaging system hardware itself, the imaging information of different object materials is not much different, which leads to the weakening of the distinguishability of the texture features and reduction of the robustness. In addition, only the texture image of the object cannot accurately reflect the object properties associated with the material. For example, when we buy clothes online, we can only purchase them according to the picture information of the objects, and we cannot accurately feel the difference in the material of the clothes. Tactile sense is a way of identifying the material of objects in people's daily life. The information contained in it can make people feel the material properties of different objects more intuitively [2]. Introducing touch into the field of object recognition is the future development trend, which can make people use more sensory information to judge and recognise the material properties of objects, which has important research significance.

Chang Liu et al. proposed to improve the knowledge of human brain is the key to improve artificial intelligence, and multi-modal learning is the key [3]. This paper studies a visual-tactile cross-modal learning method.

For texture images of different objects, the haptic feedback signals under the corresponding tool interactions can make people better sense their tactile characteristics. Research on haptic modelling has been around for decades, and in haptic modelling it is usually the state of the tool (such as the speed of the tool), the state of the textured surface (such as the properties of the texture), and the output is the vibrational haptic signal. The introduction of tactile signals into object recognition can enrich the properties of the material of the object being sensed, and have a great value in various fields, such as medicine and engineering [3].

2 Related work

Currently, the perception of the tactile properties of objects depends mainly on the display of hardware. Komura et al. studied a tactile mouse that converts the design parameters of objects into corresponding bump patterns for the palm of the user [4]. The principle is to connect the mouse to the dot matrix convex plate and move the mouse cursor on the image. While moving the mouse cursor towards the virtual object on the screen image, the corresponding touch will be generated to stimulate the palm of the subject. The disadvantage is that the stimulating touch of the dot matrix raised plate feedback is limited, so it is only suitable for the recognition of distinct object image shapes. Strese et al. developed a tactile mouse that can display the material properties of an object [5]. It mainly relies on the sensor hardware's judgment of the local data set and the real-time interaction is poor. The latest research in the modelling of haptic data focuses on the real-time interactivity of tool states. Previous research has mainly used the nature of vibration to describe the normal force and velocity of the recording tool [6, 7], which are encoded in autoregressive models. Their model successfully mapped the state and vibration mode of the tool. However, a single model generates a vibration signal that only supports a single texture used during training. Therefore, when attempting to generate vibration tactile of another texture, it is necessary to replace the model with another one. Obviously, the universality of the model is not sufficient.

Due to the complex mapping between the input and output of the tactile signal and the limitations of the level of the experimental tool, it is sometimes difficult to simultaneously take both the state of the tool and the state of the surface of the texture image as input. At present, there is no experimental model that can feedback its corresponding tactile signal very well. In the past two years, the use of generative adversarial networks (GANs) to generate new samples from high-dimensional data distribution (such as images) has been widely used, and is showing promising results in synthesising real-world images [8–10]. Previous studies have shown that GAN can efficiently generate images on labels [11], text [12], and so on. Ujiitoko et al. [9] used GAN to simulate time series data distribution, which indirectly convert vibrational tactile signals into images, and finally generate vibrato tactile signals dependent on texture images or texture properties. Although the model realises the tactile transformation of texture images, the evaluation of the tactile generation results of this model still remains on the artificial subjective perception discrimination, and
there is no objective quantitative evaluation system, which makes the research results not a good criterion for judging.

Based on the deep learning, this paper takes the texture attribute of the image as input, and simulates the tactile signal of the image texture through training and learning, and realizes the cross-modal transformation of the image texture to the tactile signal. For the tactile signal generated by the model, in addition to the artificial subjective evaluation, this study also samples the vibration frequency of the tactile signal, draws the waveform diagram, and compares the drawn waveform diagram with the frequency sampling diagram of the original real tactile signal. The similarity of the waveform is invoked as a criterion for the final experimental results, thus providing a quantitative criterion.

3 Model description

The hardware device and the overall model established in this study are shown in Fig. 1.

The model is roughly divided into the following three parts:

(i) Image feature extraction classification: The image classifier is trained and the texture image is encoded into a tag vector C. The learning network used in this part is resnet-50 [13]. After the data set is substituted into the training, the final accuracy layer of resnet-50 is removed, and the vector with image label information is defined in the penultimate layer as C.

(ii) Transform features into a spectrogram: The GAN is trained, and finally the generator is used by the GAN to generate spectrograms, which is the representation of the vibration signal in the time–frequency domain. The GAN used in this part is deep convolutional GAN (DCGAN) [14], and its input is C and the noise signal is Z from the previous step. After the training is expected to be completed, the discriminator of the network is removed, and the generator is retained.

(iii) Convert into a tactile signal: The spectrogram generated by the generator in the preceding step is converted to the tactile signal using the Griffin–Lim algorithm [15], and is directly fed back to the palm through the mouse.

4 Converting texture image into a haptic signal

The data set used throughout this study is LMT-108-Surface Materials-Database [16]. There are 108 different texture images in the data set. According to the difference in material, it is divided into nine categories. The data set selected by this study is shown in Fig. 2.

The data set also includes various types of texture images and acceleration signals generated by the corresponding tools for sliding tasks. Each type has ten sets of acceleration signal samples, and each set contains three time series signals of X, Y, and Z.

4.1 Extraction of image features

(i) Pre-processing: Since the sample of the original data is small (only ten pictures per class), in the training stage, according to the principle of data enhancement, the original data is mirrored, rotated, scaled, randomly cropped etc. The sample size was amplified to >200, the total sample size was 2100, and the size of each picture was 128 × 128.

(ii) Training: In the process of extracting and classifying image features, if the conventional training network is a shallow network, the recognition efficiency of the network cannot be significantly improved, but the deeper the network layer, the more the gradient disappears. Obviously, the training effect of the network will not be very good. The Resnet network is constructed using the skip connection residual module [13] so that it can train to get a deeper network. In this step, the training network used in this study is the deep residual network resnet-50.

At the same time, the network optimisation method adopted in this step network is Adam optimisation with a minimum batch size of six. The Adam algorithm is an algorithm for first-order gradient-based optimisation of stochastic objective functions, based on adaptive estimates of lower-order moments. The method is simple and efficient, is invariant to the diagonal rescaling of the gradients, and is well suited for problems that are large in terms of data or parameters [17, 18]. Some other specific parameter settings for this classification model network are shown in Table 1.

The final training accuracy in the test data set is 0.98421 and loss is 0.0821. The final output layer of the model is vector C with image label information. As the data in this study is divided into nine categories, C is a nine-dimensional vector.

4.2 Transforming features into spectrograms

GAN is inspired by the two-person, zero-sum game theory [8]. Two players in the GAN model are the generative model and the discriminative model. The generation model G is used to learn the distribution of the sample data, and the discrimination model D is used to discriminate the true and false probability of generating the output data of the model G. The game training through G and D can make G finally generate the image information we want.

The purpose of this study is to create a tactile signal, which is a time series of data. Although GAN is good at generating 2D...
images, but cannot generate time series data [19], so this paper selects the spectrogram as the representation of the acceleration signal, so as to train GAN to generate the spectrogram, which is the cause of this step.

Before using the GANs to automatically get the spectrogram we want, we need to use the acceleration signal of the original database to get the reference sample we want (real spectrograms). The acceleration signal is 4 s long and the sampling rate is 10 kHz. A short-time Fourier transform (STFT) solution is required to perform this step. STFT is a mathematical transformation related to the Fourier transform that determines the frequency and phase of a local region sine wave of a time-varying signal [20]. The STFT-processed signal has localised characteristics in the time and the frequency domains [21]. In this step, the result signal of the short-time Fourier transform is subjected to logarithmization and normalization, and information within the 0–1.625 time domain and the 0–256 Hz frequency domain is extracted, and finally a spectrogram is generated based on the information. The parameters of other STFTs are shown in Table 2.

The converted result is shown in Fig. 3 (G1–G9 from top left to bottom right).

The above operation only obtains the training data, and the purpose of this step is to autonomously generate the desired spectrogram with reference to the training data. The GAN structure we refer to here is DCGAN [14]. DCGAN is a good improvement after GAN, greatly improving the stability of GAN training and the quality of the generated results. However, it is limited to only one type of training. Although AC-GAN [22] can resolve this problem, AC-GAN training is cumbersome, and the training results are greatly interfered by the inter-class and are unstable. It is usually used for multi-feature learning of similar images [23]. The main purpose of this study is to complete the texture image-tactile cross-modal conversion, which requires high stability for the generated results. Therefore, we combined the idea of ensemble learning with DCGAN to build an ensembled GAN, which makes the model more robust while training in multiple categories.

In this study, for each training class, each category is trained separately. This has the advantage of ensuring the diversity within the class and ensuring the differences between classes. The control variable used to determine the generator is vector C with label information in Section 4.1. The brief network structure of ensembled GANs utilised for training in this study is shown in Fig. 4.

The network is mainly composed of two parts, discriminator D and a generator G. The input of the generator is composed of C and Z. C is the image vector of the label information obtained in the previous step (in Section 4.1), and Z is the random noise. In this study, the dimension of the Z is set to 100. The discriminator D first trains the spectrogram data set, establishes a judging index system, then takes the spectrogram obtained by the generator G as an input, and finally completes the discrimination and classification of the generated spectrogram by D. After 20 epochs of confrontation training, the result is compared with the original. Here, three categories are selected for comparison. As shown in Fig. 5, the results generated by the model are very similar to the original data.

### 4.3 Converting into a tactile signal

In this step, the Griffin–Lim algorithm is used to generate the corresponding audio signal from the spectrogram generated by G in the previous step. The Griffin–Lim algorithm can reconstruct the signal well according to the spectrum [15], and the audio signal can be generated by the power amplifier signal. After visualising the signals, each type of the corresponding signal oscillating is shown in Fig. 6, and it can be observed that there is a significant difference between the categories.

### 5 Result analysis

#### 5.1 User test

Tactile rendition devices are generally stimulating the tactile receptors of the skin by various methods, such as air bellows or nozzles, vibrations generated by electrical excitation, microneedle arrays, direct current pulses, and functional neuromuscular stimulators. The tactile output module of this study is an audio vibrator, which belongs to the vibration device generated by electric excitation. Its advantage is that the motor is used as an actuator, which can obtain torque in any direction with relatively fast response speed, and is very sensitive to different frequencies. The device can be used to simulate the tool interaction state of the original haptic signal acquisition, which is the main reason for selecting this device as the haptic output. Its disadvantage is that because of the excessive sensitivity to the signal, it may sometime

| Table 2 | STFT parameters | Parameters | Values |
|---------|----------------|------------|--------|
| Fs      | 10,000         | hamming window | 512   |
| noverlap| 128            | number of points in fast fourier transform | 512   |
| time    | 1.625 s        | frequency region | 0–256 Hz |

**Fig. 3** Raw data spectrogram

**Fig. 4** Ensembled GANs network structure

**Fig. 5** Comparison of results
cause instability of the system, thereby damaging the effect of
force tactile reproduction. Other devices, such as DC pulses and
functional neuromuscular stimulators, are also suitable for this
study, but they have shortcomings such as insufficient stability and
high cost. Based on the above considerations, we finally chose the
audio vibrator as the tactile output module.

This study first carries out an artificial subjective test. The test
equipment and process used include: the output of the PC through
the power amplifier and the tactile output module is nested inside
the wireless Bluetooth mouse. While testing, the mouse clicks on
different test images, different tactile feedback can be felt on the
palm of a person, specifically, as shown in Fig. 7.

There were 50 users (25 male and 25 female individuals) tested
for this study, aged between 20 and 30, and the senses of these
people were normal. The test is mainly divided into two parts: first,
shielding the visual information of the tested person performing the
output guessing test of the tactile signal and then removing the
visual mask so that the tested person can evaluate the tactile output
results of different texture images according to the visual
information. The test results are shown in Fig. 8.

5.2 Experimental data analysis

In addition to the artificial subjective evaluation, this study also
compares the data obtained from modelling the original tactile
signals and the tactile signals generated by the model learning, and
finally quantified and analysed the waveforms of the two as shown
in Fig. 9.

Specific calculation methods, such as metal net, are shown in
Fig. 10.

The metal net 1 is a frequency sampling waveform of the
original haptic signal, and the coordinates of the centre line are
obtained as $(x_1, y_1)$; the metal net 2 is a waveform diagram drawn
by signal sampling generated by the model of this study, and the
coordinates of its centreline are obtained as $(x_2, y_2)$. Here, $x$ is the
time domain information, $y$ is the frequency domain information,
and $x_u$ is equivalent to $x_2$. By comparing the frequency difference
of each point at the same time domain, the overall similarity
between the two can be obtained.

The calculation of similarity is

$$ S = \frac{1}{e^{\frac{1}{2}(x_1-x_2)^2 + (y_1-y_2)^2}}. $$

The index of the denominator in the above formula actually finds
the Euclidean distance between the generated data waveform and
the original data waveform at the same time domain point, so that
the waveform difference between the two in the same time domain
can be obtained. The denominator has a value range of $[1, +\infty)$, so
the range of similarity $S$ values is $[0,1]$. When the frequency
difference between the two is larger, the closer $S$ is to 0, the smaller
the frequency difference between the two, the closer $S$ is to 1. In
this way, the similarity between the two waveforms can be judged,
so that the quality of the generated haptic signal can be judged.
Finally, the evaluation results of the nine experimental data sets of
this study under this calculation method are shown in Fig. 11.

![Fig. 6 Tactile signal waveform display](image6.png)

![Fig. 7 Hardware diagram](image7.png)

![Fig. 8 User experience result statistics](image8.png)

![Fig. 9 Left picture shows the original data waveform and right picture shows the generated data waveform](image9.png)

![Fig. 10 Signal comparison](image10.png)
This work mainly studies the visual-tactile cross-modal transformation based on deep learning. First, the Resnet is used to obtain the classification information of the image. The next step is to use ensemble GANs to obtain the spectrogram generator G, and then combine the classification information of the resent output with G. It is possible to automatically generate spectrograms of different texture images, and then use Griffin-Lim algorithm to convert the spectrogram into a tactile signal. Finally, haptic signals of different texture images are fed back to the palm through the Bluetooth mouse. The experimental results show that the model can transform the visual information of the texture image into tactile information, which is close to the actual tactile sensation. At the same time, the experimental results also scientifically verify the tactile evaluation method in this study.

Of course, the research work still has limitations. In the actual tactile test, the tactile feedback of the object material is constrained by the visual classification effect and the visual error will be continued into the tactile feedback. Despite this, this research has great application potential.

7 Acknowledgment
This work was supported in part by the National Natural Science Foundation of China under Grant 61673238.

8 References
[1] PRICAI 2018: ‘Trends in artificial intelligence’, 2018 Proc. 15th Pacific Rim Int. Conf. on Artificial Intelligence, Nanjing, China, 28–31 August 2018

Fig. 11 Evaluation of results

6 Conclusion
This work mainly studies the visual-tactile cross-modal transformation based on deep learning. First, the Resnet is used to obtain the classification information of the image. The next step is to use ensemble GANs to obtain the spectrogram generator G, and then combine the classification information of the resent output with G. It is possible to automatically generate spectrograms of different texture images, and then use Griffin-Lim algorithm to convert the spectrogram into a tactile signal. Finally, haptic signals of different texture images are fed back to the palm through the Bluetooth mouse. The experimental results show that the model can transform the visual information of the texture image into tactile information, which is close to the actual tactile sensation. At the same time, the experimental results also scientifically verify the tactile evaluation method in this study.

Of course, the research work still has limitations. In the actual tactile test, the tactile feedback of the object material is constrained by the visual classification effect and the visual error will be continued into the tactile feedback. Despite this, this research has great application potential.

7 Acknowledgment
This work was supported in part by the National Natural Science Foundation of China under Grant 61673238.

8 References
[1] PRICAI 2018: ‘Trends in artificial intelligence’, 2018 Proc. 15th Pacific Rim Int. Conf. on Artificial Intelligence, Nanjing, China, 28–31 August 2018

Fig. 11 Evaluation of results

6 Conclusion
This work mainly studies the visual-tactile cross-modal transformation based on deep learning. First, the Resnet is used to obtain the classification information of the image. The next step is to use ensemble GANs to obtain the spectrogram generator G, and then combine the classification information of the resent output with G. It is possible to automatically generate spectrograms of different texture images, and then use Griffin-Lim algorithm to convert the spectrogram into a tactile signal. Finally, haptic signals of different texture images are fed back to the palm through the Bluetooth mouse. The experimental results show that the model can transform the visual information of the texture image into tactile information, which is close to the actual tactile sensation. At the same time, the experimental results also scientifically verify the tactile evaluation method in this study.

Of course, the research work still has limitations. In the actual tactile test, the tactile feedback of the object material is constrained by the visual classification effect and the visual error will be continued into the tactile feedback. Despite this, this research has great application potential.

7 Acknowledgment
This work was supported in part by the National Natural Science Foundation of China under Grant 61673238.

8 References
[1] PRICAI 2018: ‘Trends in artificial intelligence’, 2018 Proc. 15th Pacific Rim Int. Conf. on Artificial Intelligence, Nanjing, China, 28–31 August 2018

Fig. 11 Evaluation of results

6 Conclusion
This work mainly studies the visual-tactile cross-modal transformation based on deep learning. First, the Resnet is used to obtain the classification information of the image. The next step is to use ensemble GANs to obtain the spectrogram generator G, and then combine the classification information of the resent output with G. It is possible to automatically generate spectrograms of different texture images, and then use Griffin-Lim algorithm to convert the spectrogram into a tactile signal. Finally, haptic signals of different texture images are fed back to the palm through the Bluetooth mouse. The experimental results show that the model can transform the visual information of the texture image into tactile information, which is close to the actual tactile sensation. At the same time, the experimental results also scientifically verify the tactile evaluation method in this study.

Of course, the research work still has limitations. In the actual tactile test, the tactile feedback of the object material is constrained by the visual classification effect and the visual error will be continued into the tactile feedback. Despite this, this research has great application potential.

7 Acknowledgment
This work was supported in part by the National Natural Science Foundation of China under Grant 61673238.

8 References
[1] PRICAI 2018: ‘Trends in artificial intelligence’, 2018 Proc. 15th Pacific Rim Int. Conf. on Artificial Intelligence, Nanjing, China, 28–31 August 2018