INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 19, Issue: 5, Month: October, Year: 2024
Article Number: 6457, https://doi.org/10.15837/ijccc.2024.5.6457

Deep Multimodal Fusion of Visual and Auditory Features for Robust Material Recognition

Yifei Shi, Huei Ruey Ong, Shuai Yang, Yuxin Fan

Yifei Shi
1.Geely University of China
2.DRB-HICOM University of Automotive Malaysia
shiyifei@guc.edu.cn

Huei Ruey Ong*
DRB-HICOM University of Automotive Malaysia
*Corresponding author: hueiruey@dhu.edu.my

Shuai Yang
Geely University of China
yshuai@digcow.com

Yuxin Fan
Geely University of China
yxfan@digcow.com

Abstract

This paper presents a deep neural network incorporating visual and auditory data fusion to enhance material recognition performance. Traditional recognition techniques relying on single data modalities face accuracy and robustness limitations, especially in complex real-world environments. To address these challenges, we develop a multimodal fusion-based model. The proposed approach first extracts features from input images and sounds separately using CNNs and spectral analysis. A concatenation layer then integrates the visual and auditory features. Extensive experiments demonstrate superior material classification over uni-modal methods, with 100% test accuracy across seven material types. The multi-modal fusion model also demonstrates stronger resilience to noise and illumination variations. This research provides a valuable foundation for robust material perception in intelligent systems.

Keywords: material recognition, deep neural network, visual information, auditory information, feature fusion.

1 Introduction

In recent years, the field of material recognition has made significant advances, yet the accuracy and robustness of unimodal recognition systems—relying solely on either visual or auditory data—remain limited, particularly in complex and variable environments. This study aims to address these limitations by developing a multimodal deep neural network framework that integrates both visual and auditory information, enhancing both the precision and reliability of material recognition.

In real-world scenarios, objects are usually represented in multiple ways, with images and audio each having the ability to independently characterize the information about the object, but at the same time having limitations and shortcomings. Visual information is mainly formed by the reflection of light in different wavelengths, which is then captured by the light-sensitive elements, thus presenting a visual representation of the object. Visual representations are vivid and comprehensive, covering multiple information dimensions such as color, brightness, and shape. The main reason why visual information has been used for material recognition for a long time is that visual information often contains the intuitive characteristics of the target object, so it can synthesize various types of features to describe and judge the target object as a whole [1] In practical applications, noise and light are the most important factors in material recognition. In practice, factors such as noise and changes in illumination may interfere with the visual information, thus affecting the accurate recognition of the material of the object [2]. In the process of image acquisition, the equipment may not be able to recognize the material of the object. During the image acquisition process, noise may appear in the image due to various reasons such as equipment performance and environmental conditions. These noises may be manifested as random pixel point variations, color distortion, etc. in the image. For object material recognition, noise may lead to greater interference in the spectrogram of the audio information related to the object. Changes in light brightness may be caused by factors such as changes in the natural environment, differences in the performance of lighting equipment, etc. When the light brightness changes, the reflective properties of the surface of the object also change, thus affecting the image of the object [3, 4]. In contrast, audio information consists of sound waves that convey specific properties of objects through acoustic signals. Audio information, corresponding to acoustic waves, can enables detection of an object's position and distance, and is insensitive to obstacles and other disturbing factors, and can still effectively transmit information, which has an advantage over visual images in this regard. When we touch an object, we not only perceive its texture by touch, but also perceive its sound by hearing, which is often closely related to the material of the object. However, audio information is not intuitive and requires the design of complex models to compute and simulate the perceptual properties of the human ear, which can have a large impact on the task, especially in the presence of noise interference. Auditory information is subject to a certain degree of interference in both its time and frequency domains with ambient noise. White noise causes the original sound in the time domain to be "drowned out" at low signal-to-noise ratios, whereas the original sound in the frequency domain, although preserved, is also affected. Traditional material recognition methods have predominantly focused on unimodal approaches. Visual-based techniques, while rich in detail, often falter under poor lighting conditions or when objects present visually similar textures. Auditory methods, conversely, can offer additional context through sound analysis, such as tapping responses, which are less susceptible to visual obstructions but can struggle with background noise interference. These inherent limitations highlight a crucial gap in current recognition technologies' ability to adapt to diverse and real-world scenarios.

The integration of multimodal information presents a promising avenue to surmount these challenges. By fusing visual and auditory data, the proposed neural network aims to leverage the strengths of both modalities, thus significantly reducing the dependency on a single type of sensory input and increasing the system's robustness against environmental variabilities. Previous studies have shown that multimodal systems can achieve higher accuracy than their unimodal counterparts by providing a richer representation of objects under study. In order to overcome the above limitations and improve the accuracy of material recognition, this study proposes a deep neural network that fuses visual and auditory information. The network first preprocesses the input image and audio signals, then extracts the image features and audio features respectively, and fuses the extracted features to finally construct a deep neural network that fuses visual and auditory information to recognize materials, as shown in Figure 1. Specifically, the network first preprocesses the input image to extract features such as color, texture, and shape of the image. Then, the network preprocesses the input audio signal to extract the spectral features or sound characteristics of the audio signal. Next, the network fuses the extracted image features and audio features to generate a feature vector that incorporates visual and auditory information. Finally, the network uses a deep neural network to classify and recognize the fused feature vector, outputting the corresponding material labels. Thus, after a sufficient number of training sessions, the network can gradually learn how to accurately map the input image and audio signals to the corresponding material labels. By fusing visual and auditory information, this deep neural network can take full advantage of both modal information to improve the accuracy of material recognition. At the same time, the network also has good robustness and can effectively resist the influence of disturbing factors such as noise and illumination changes. In addition, the network can be trained using a large amount of existing image and audio data to

further improve the performance and generalization ability of the model. Our multimodal approach specifically targets issues of accuracy in visually complex scenarios and robustness against auditory noise disturbances. For instance, the ability to accurately recognize materials in a cluttered visual scene or in a noisy environment could greatly benefit various applications, from industrial sorting to interactive educational technologies. The deep learning framework developed in this study is designed to intelligently integrate and analyze the complementary data from both visual and auditory sources, thereby addressing these critical challenges.

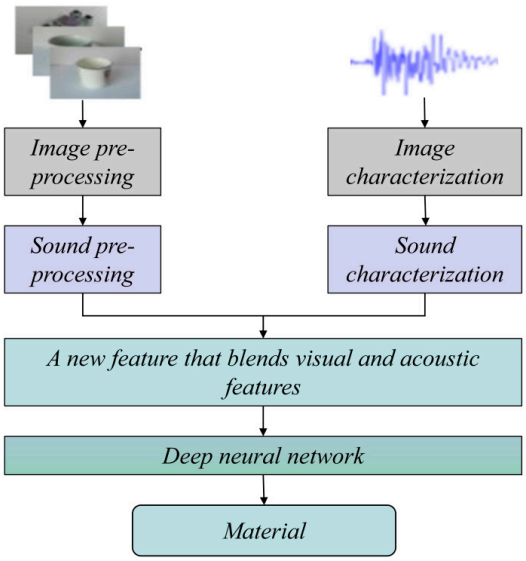

Figure 1: Deep neural network for material recognition fusing visual and auditory information

2 Literature review

In the real world, the material of an object is one of its important properties. Therefore, the accurate identification of the material is crucial when constructing virtual models [5, 6] . Datasets and learning algorithms are important tools to handle the material recognition task and provide strong support to researchers. Earlier studies mainly utilized visual information for material recognition [7, 8, 9]. However, with the continuous development of technology, data containing multi-modal information such as visual, auditory, and haptic information have gradually emerged and have received much attention in the field [10, 11, 12, 13]. The data is also used as the basis for the development of the visual, auditory and tactile information. Among the many types of information, visual and auditory ones are crucial.

In the field of computer vision, researchers have developed numerous vision-based material recognition algorithms. These algorithms commonly employ features such as color, texture, and shape of an image for material recognition, and the material associated with a particular color can be identified by analyzing the color histogram or color features of an image [14, 15]. By analyzing the texture features of an image, a material with a specific texture can be identified [16, 17]. The material with a specific shape can be identified by analyzing the shape features of the image [18]. The main reason why visual information has been used for material recognition for a long time is that visual image information often contains the intuitive characteristics of the target object, so it is possible to synthesize various types of features to describe and judge the target object as a whole [19]. The current application of auditory information in material recognition is mainly reflected in two aspects, one is based on the method of sound signal analysis, and the other is based on the method of machine learning. The former refers to the use of the characteristics of the sound signal to identify the material, specifically, it is necessary to analyze the frequency, amplitude, phase and other parameters of the sound signal, from which to extract the features that are closely related to the material of the object, so as to realize the material recognition [20, 21]. The latter refers to the use of a large amount of sound data for model training, so that the model from the sound data can automatically extract features related to material properties and their recognition [22].

Relying on information from one source alone has its shortcomings. At present, the fusion processing of visual and auditory information, which are two different modalities, has not been fully investigated. However, with the development of audio and image feature extraction techniques, audio-visual fusion techniques are gradually being developed. Studies have been conducted to provide a comprehensive overview of the 2019 Audiovisual Speaker Recognition Evaluation (SRE19), detailing its tasks, performance metrics, data, evaluation protocol, results, and system performance analysis. The evaluation consists of two main aspects: audio-only

versus audiovisual, and an optional visual-only [23] . In light of the limitations of traditional approaches in correlating audio signals with visual elements, especially in complex scenes with similar musical instruments, the research introduced "musical gestures", a structured representation based on keypoints designed to explicitly model the musician's body and finger movements during a performance. Their approach integrates a network of context-aware graphs for combining visual context with body dynamics, and then links these movements to the corresponding audio signals through an audio-visual fusion model [24]. The researchers proposed a novel approach of integrating attentional fusion blocks during the coding of audiovisual data, which bridges the gap of existing studies between different modal switches by exploiting the correlation between the two modalities to enrich the audiovisual representation [25]. In addition, a new method for detecting longitudinal tears in conveyor belts, called audiovisual fusion (AVF), which goes beyond the limitations of purely visual detection and provides higher accuracy and reliability in identifying conveyor belt damage [26]. The researchers proposed an audio-visual fusion model that combines deep learning features with a mixed brain affective learning (MoBEL) model inspired by the limbic system of the brain, aiming to simultaneously learn from the combined audio-visual features to recognize the spatio-temporal correlations inherent in the video [27]. The audiovisual fusion multi-stage cross-attention technique utilizes multimodal representations for weakly-supervised action or event localization in unedited videos, which effectively improves the accuracy in video recognition [28]. The ability of auditory and visual detection of sound sources has been exploited and a multi-source localization method using neural networks with audio and visual signals has been proposed [29]. Dimensional emotion recognition from video using fusion of facial and sound modalities has highlighted the shortcomings of existing fusion techniques that rely heavily on recurrent networks or traditional attention mechanisms and fail to effectively utilize the complementary nature of audiovisual modalities, and technological innovations based on this [30]. AV-SAM is an audiovisual localization and segmentation framework based on Segment Anything Model (SAM). The framework aims to efficiently generate masks for sound objects and associate them with audio in various tasks [31]. Based on the machine learning depth approach, four dynamic neural network architecture-based methods for single-peak object recognition in colleges and universities are proposed in the study considering performance issues and stability requirements, and the results of the study show that the top-1 error rate of the single-frame tactile model is reduced by 78% and the average accuracy is improved by a factor of 2.19 by augmenting the tactile-based visual information using different strategies [32]. Based on the existing multi-scale and multi-directional architecture of the bionic model, the researchers constructed the Mel frequency cepstrum coefficients on the basis of the audio code to realize the bimodal recognition with the fusion of visual and auditory perceptions, which effectively improves the recognition efficiency of the model [33]. The abstract visual representation of the amplitude envelope cues of the target sentence is beneficial to speech perception in complex listening environments, providing a research reference for speech sensory recognition in noisy environments [34]. The speed of data information acquisition is improved by generating representative vectors through CNN-based recursive modeling and fusion module, which leads to a substantial improvement of the overall performance in audiovisual fusion [35].

Considering that existing studies still have great difficulties in object material recognition, this study proposes a visual and auditory fusion method for object material recognition, and validates the effectiveness of the visual and auditory fusion method under the multimodal fusion perspective and combined with self-constructed library data.

3 Research methodology

3.1 Database of visual and auditory information features

In this study, aimed at advancing object material recognition, a high-resolution video camera (720p×1280p, 30 frames/sec) was employed to capture high-quality images against specifically designed backdrops. To ensure robust generalization of the carefully trained model for classifying materials, the video data spans a variety of lighting conditions—including natural, green, red, and blue light—and accounts for differences in shooting angles and object sizes and shapes. This approach enhances the model's exposure to diverse scenarios, contributing to a comprehensive training dataset. The experiment involved seven categories of materials: wood, plastic, steel, profiles, cotton fabrics, ceramics, and glass. Each category was meticulously documented under varied lighting and angular conditions to maintain data diversity and validity. Subsequently, a specialized image processing tool was used to select and compile the desired images into a database, resulting in 10,870 training samples and 4,860 test samples.

During the sound data collection phase, we carefully selected and analyzed sounds produced by colliding or knocking on seven commonly used materials in daily life, including plastic, ceramics, silk fabric, wood, aluminum alloy profiles, metal, and glass. The acoustic characteristics of these materials can vary significantly depending on their composition and shape, introducing variability into the recorded sound signals. To control environmental noise, the recording sessions were conducted in a laboratory setting during early morning hours.

A standard 2B pencil was chosen for its consistent hardness, which minimizes variability in the produced sound when tapping different materials. This consistency is crucial for comparing the acoustic responses of diverse materials under controlled experimental conditions. To further ensure the reliability and diversity of the sound dataset, three distinct samples from each material category, varying in size and specification, were audibly tested. This rigorous methodological framework supports the creation of a rich and representative dataset, providing a solid foundation for subsequent analyses and model training.

3.2 Visual information feature extraction

Convolutional Neural Networks (CNNs) were selected for visual feature extraction due to their proven effectiveness in handling spatial hierarchies in images, which is crucial for distinguishing material textures and patterns.

The basic structure of a convolutional neural network (CNN) is shown in Figure 2. After the input data is convolved in the first layer, shallow features are obtained and the feature map is output. Next, the desired features are selected through the pooling layer to achieve compression and a new feature map is constructed. Subsequent convolution and pooling operations similarly process the output of the first layer to gradually obtain deep features. Finally, the fully connected layer accomplishes the corresponding tasks based on the acquired deep features. If the fully connected layer is removed and multiple convolutional and sampling layers are superimposed, a multilayer CNN model can be constructed for deep feature extraction.

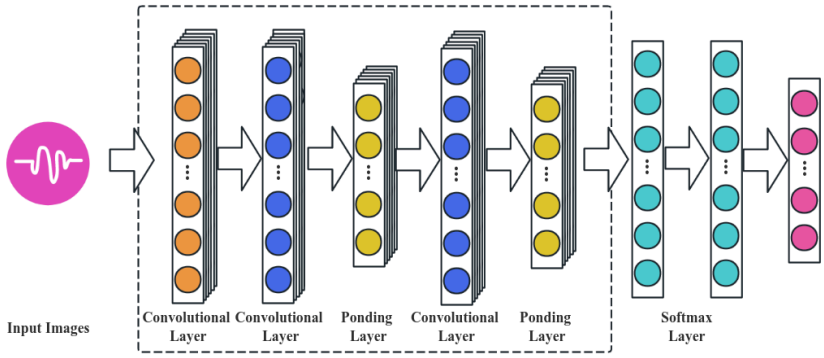

Figure 2: Basic structure of CNN model

Following the above strategy of CNN model for feature extraction, we can choose the pre-trained VGG19 model and remove the fully-connected layer at the end of it, and use it as a component for image information feature extraction.The VGG19 model is essentially a multilayer CNN, which is a deep model based on CNNs and contains 19 layers, including 16 convolutional layers and 3 fully-connected layers. Although the pre-trained VGG19 model is effective for image feature extraction, there is a risk of overfitting to the ImageNet dataset rather than to our specific materials dataset. To mitigate this risk, further fine-tuning on material-specific images was conducted to ensure the model better generalizes to the characteristics of different materials. In practice, there is no significant difference between VGG19 and VGG16. For VGG16, it achieves a certain size of sensory field (the local information of the input picture that affects the picture output) by replacing the original large-size convolutional kernels (11×11, 7×7, 5×5) with multiple 3×3 convolutional kernels. The advantage of this change is that several small convolutional kernels outperform a single large convolutional kernel while keeping the size of the receptive field unchanged, because in a multilayer network, multiple convolutional operations increase the depth of the network, which in turn results in a more complex descriptive capability.

In addition, the advantage of small convolutional kernels in neural networks is that they have fewer parameters and lower computational cost compared to large convolutional kernels, and thus have an important position in deep learning models. In the VGG model, we replace the large convolutional kernels with all 3×3 sized convolutional kernels, which improves the descriptive ability of the network while keeping the perceptual field of the large convolutional kernels unchanged. This move not only reduces the computation of model parameters, but also improves the performance of the neural network.The structure of the VGG model also becomes more concise, with the same size of convolutional kernels (3×3) as well as maximal pooling (2×2) uniformly used throughout. With the same basic structure as a traditional convolutional neural network (CNN), the VGG model acts as an image feature extractor after removing the last fully connected layer. That is, when the input is a colorful square image with a side length of 224, the output is 7×7×512 image features. This structure largely improves the model's ability to extract image features and lays the foundation for subsequent classification

and recognition tasks. By adopting a small convolutional kernel and maintaining a concise network structure, the VGG model achieves a good balance between neural network performance and computational efficiency. Meanwhile, when designing deep learning models, the effects of convolution kernel size and model structure on performance should be fully considered to achieve higher computational efficiency and better generalization ability.

3.3 Characterization of auditory information

As a kind of non-stationary signal, sound signal behaves as a function of sound pressure with time, and has the characteristics of time-varying, non-stationary and large dispersion. At this stage, we mainly use time domain, frequency domain and cepstrum analysis methods to explore the sound signal. In practice, time-domain and frequency-domain analysis methods are widely used in the field of sound signal analysis. Especially, the time-domain based analysis method is small in operation, easy to realize, and has a clear physical meaning. As shown in Figure 3, the results of time-domain analysis of ceramic, metal, wood and plastic sound signals are demonstrated. The short-time energy and short-time average over-zero rate of the sound signal can be calculated by setting the frame length and frame shift reasonably during the partitioning process. However, the results of a large number of studies show that the time-domain features, such as short-time energy and short-time average zero crossing rate, obtained from the time-domain analysis of the sound signals emitted by objects of different materials at the time of collision or percussion are less distinguishable. Therefore, in the classification and recognition of material sound signals, it is difficult to achieve the expected results if only the time domain features such as short-time energy and average zero crossing rate are selected as the characterization of sound signals.

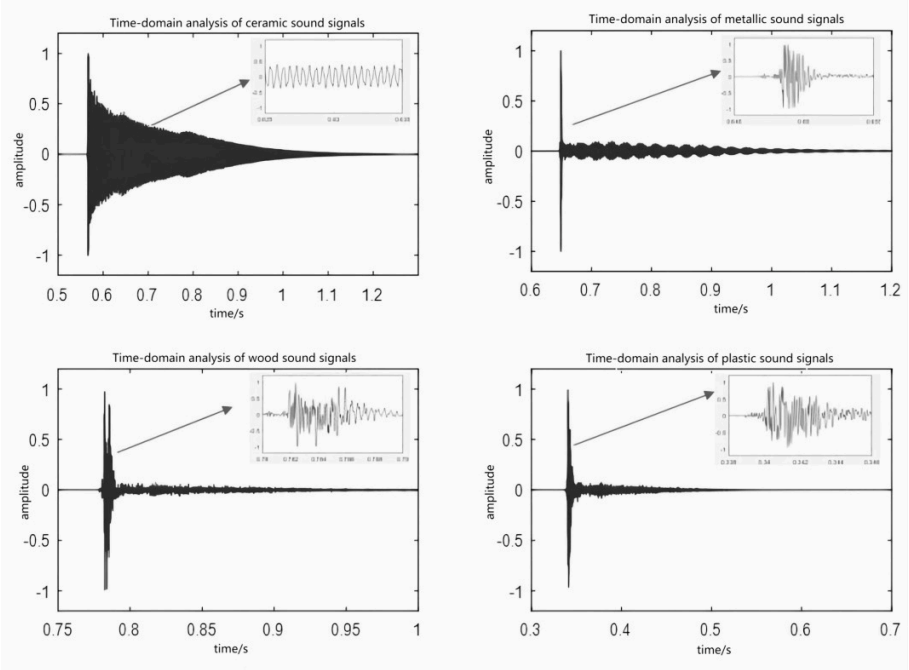

Figure 3: Time-domain analysis of sound signals from different material objects

Spectral analysis has been hailed as the most widely used and effective means of analyzing sound signals in existing research, due to the fact that the spectrum of a sound signal can reveal the characteristics of the excitation spectrum and the frequency of the vocal tract. Currently, there are many methods for speech feature extraction, and Mel Frequency Cepstrum Coefficient (MFCC), Linear Predictive Analysis (LPC), and Short-Time Fourier Transform (STFT) are commonly used. Mel Frequency Cepstral Coefficients (MFCCs) were used for auditory feature extraction because they closely mimic the human auditory system's response and are particularly effective in extracting relevant features from complex sound waves, making them ideal for distinguishing materials based on their sound upon impact.

Aiming at the non-stationary and time-varying characteristics of sound signals, the short-time Fourier transform is used for sound feature extraction. The basic concept of this transform is to process the signal through a sliding time window and apply the Fourier transform to the signal within the window to obtain the time-varying spectrum of the signal. The short-time Fourier transform becomes an effective analytical tool when exploring the short-time spectrum of a sound signal over time. Speech spectrogram has a key position in the field of speech signal analysis because it reveals the dynamic spectral properties of the speech signal. In the

spectrogram representation of a sound signal, the vertical coordinates symbolize frequency and the horizontal direction represents time. The signal strength can be described by the percentage of gray-scale information on the speech spectrogram, and the gray-scale stripes express the short-time spectrum of speech at each moment.

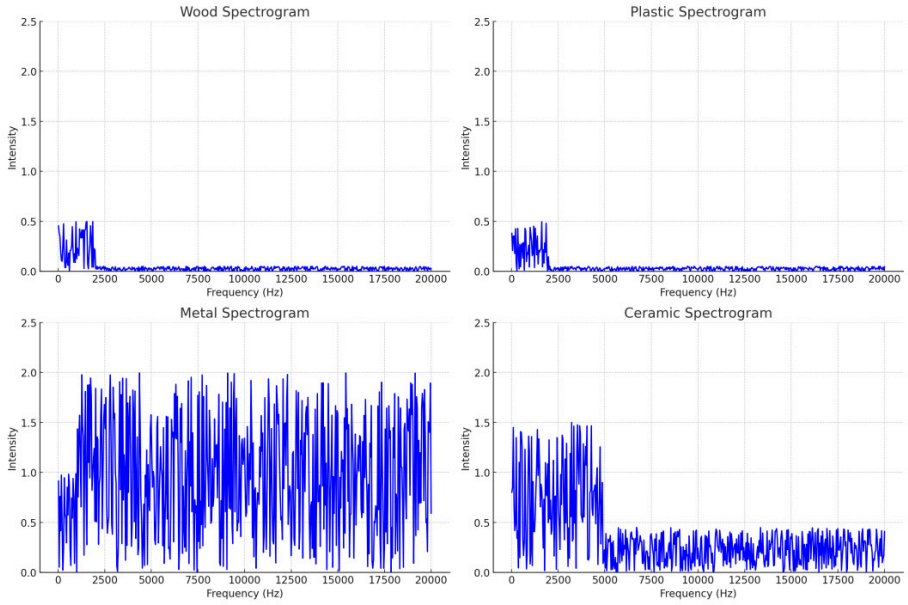

Figure 4: Results of signal spectrum analysis of different types of material objects

Figure 4 presents spectrograms of sound signals from different materials. Each spectrogram illustrates the frequency (vertical axis) and time (horizontal axis) distribution of sound, highlighting how different materials produce distinct sound signatures.

As can be observed in Figure 4, there is some similarity between the sound spectrograms produced by wood and plastic when tapping. There are fewer darker colors, which is due to the relatively low frequency of sounds produced by wood and plastic when struck. In contrast, metal produces the largest fluctuations in the graph when tapped, which is because objects made of metal emit more sound energy than the other three types of materials when bumped or tapped. In Figure 4, the order of sound energy from highest to lowest is metal, ceramic, plastic and wood, respectively. It is worth noting that the metal and ceramic spectrograms in Fig. 4 show clear horizontal lines from left to right, which represent the resonance peaks characteristic of the sound signal. Observed from the bottom to the top, the colors of the spectrograms change from dark to light, which implies that the sounds emitted by different material objects contain rich harmonic information. It can be seen that the sounds emitted by different material objects during tapping have large similarity in their frequency distributions, although they have unique signals, which increases the difficulty of sound recognition. Therefore, it is necessary to adopt deeper and more characterizing features for sound signal recognition.

In this study, Mel frequency cepstrum coefficient (MFCC) is used to extract sound signal features based on the sound fluctuation characteristics of various types of materials. For auditory feature extraction, we utilized Mel Frequency Cepstral Coefficients (MFCC) and spectral analysis. The parameters for MFCC were carefully selected: a window size of 25ms and a hop size of 10ms were used to balance temporal resolution and computational efficiency. These settings are crucial for capturing the nuanced acoustic properties of materials, facilitating the robust discrimination of their types. The Mel frequency cepstrum coefficient (MFCC) has been proven to be widely used in the field of speech recognition and plays a key role. Based on the concept of homomorphic processing, the study calculates the MFCC and introduces the Mel filter with reference to the characteristics of the human auditory system. The resonance peak is one of the key features of the sound signal, and what connects the resonance peaks is the spectral envelope of the sound signal, i.e. a smooth curve. By extracting the spectral envelope information, the researcher obtains the information that expresses the position and change of the resonance peaks, and thus proposes the deconvolution technique. In homomorphic processing calculations, the excitation source is separated from the impulse response of the sound channel and then analyzed one by one, and this homomorphic processing calculation is called a method of "nonparametric deconvolution".

In the homomorphic processing, the input is the convolution result of the sound gate excitation signal x_1(n) and the channel impulse response x_2(n) ; D[x(t)] is the first system whose role is to do additive operation on the input signal; x(t) denotes the audio signal in the time domain. The final output signal is presented in the complex cepstrum frequency domain. The convergence domains of the signals x(n) and $\hat{x}(n)$ are in the unit circle, and $\hat{x}(n)$ is the complex spectrum of x(n) , so the first D[x(t)] system can be expressed by Equation (1).

$$D[x(t)] = \begin{cases} FT[x(n)] = FT\left[x_1(n)^* x_2(n)\right] = X_1(w)^* X_2(w) = X(w) \\ \mathrm{In}[X(w)] = \mathrm{In}\, X_1(w) + \mathrm{In}\, X_2(w) = \hat{X}_1(w) + \hat{X}_2(w) = \hat{X}(w) \\ FT^{-1}[\hat{X}(w)] = FT^{-1}\left[\hat{X}_1(w) + \hat{X}_2(w)\right] = \hat{x}_1(n) + \hat{x}_2(n) = \hat{x}(n) \end{cases} \tag{1}$$

The first system D[x(t)] operates the results of the input convolution run on the inverted spectral domain $\hat{x}(n) = \hat{x}_1(n) + \hat{x}_2(n)$, therefore, the completion of the deconvolution operation brings significant advantages for the analysis and recognition of sound signals.

Based on the auditory perception mechanism of the human ear, the researchers simulated the human auditory system and designed the Mel filter. The function of this filter is to convert the conventional frequency to Mel frequency, which is beneficial to sound recognition, improves recognition accuracy and performs better. The conversion formula is as follows:

$$\mathrm{mel}(f) = 2595 * \log_{10}(1 + f/700) \tag{2}$$

In Mel spectral envelope extraction, first of all, the preprocessing step is completed for the input speech signal by fast Fourier transform. According to the study, the envelope of the sound signal belongs to the low-frequency signal characteristics, so the low-pass filtering principle can be used to filter the envelope signal. When sound signals are analyzed spectrally, the main operations include Fourier transform (FT) and Fourier inverse transform (IFT). However, in cepstrum studies, the basic principle is to perform the Discrete Cosine Transform (DCT) operation.The DCT operation removes the correlation between the sound signals and at the same time realizes the dimensionality reduction of the signal. The DCT processed signal eliminates the correlation between the dimensional signals, and the 2nd to 13th coefficients after DCT are selected for the final extraction of the MFCC features.The content of the MFCC feature extraction is shown in Figure 5.

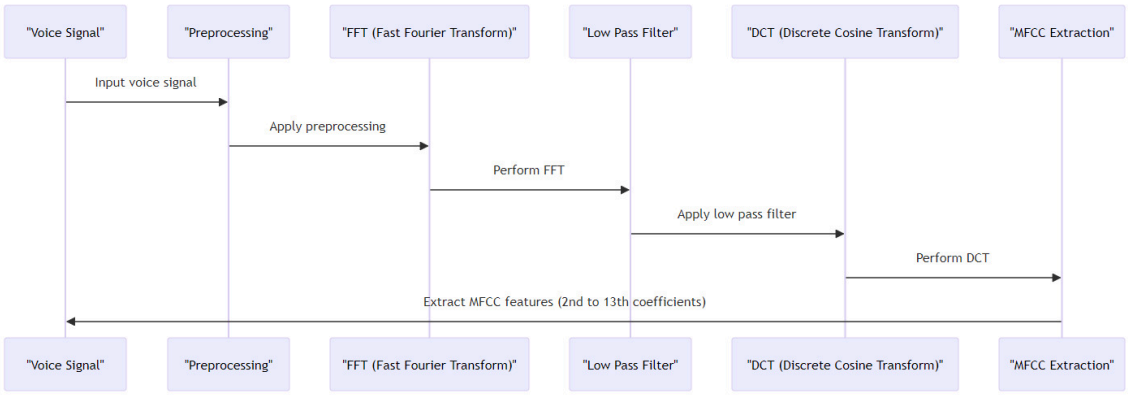

Figure 5: MFCC extraction timing diagram

In the MFCC feature extraction process, a series of preprocessing operations, including frame-splitting, windowing, and pre-emphasis, need to be performed on the input sound signal $x_i(n)$. The principle of frame splitting is to multiply the sound signal with a window function of finite length, which can be described by equation (3). In the field of speech signal analysis, the most widely used window functions include: Hamming window, Hanning window, and Blackman window, of which the expression of Hamming window is shown in equation (4).

$$y(n) = \sum_{n=-\infty}^{\infty} x(m)w(n-m) \tag{3}$$

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/(L-1)) & 0 \le n \le L-1 \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

$x_i(m)$ is the pre-processed signal and then the Fast Fourier Transform operation is performed on it, i.e. $X(i,k) = FFT\left[x_i(m)\right]$, after which the transformed signal is filtered in the frequency domain using a Mel

filter to obtain the Mel spectrum. After obtaining the Mel spectrum the calculation of the spectral line energy $E(i,k) = \left[X(i,k)^2\right]$ is done and finally the energy passing through the Mel filter S(i, m) is calculated. Calculating the energy passing through the filter S(i, m) in the frequency domain is equivalent to multiplying and adding the frequency domain responses of the Mel filter H_m(k) and E(i, k) , which can be expressed by equation (5).

$$S(i,m) = \sum_{k=0}^{N-1} E(i,k) H_m(k), 0 \le m \le M \tag{5}$$

Calculating the DCT after de-logging S(i, m) gives the MFCC:

$$mfcc(i,m) = \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} \log[S(i,m)] \cos\left(\frac{\pi n(2m-1)}{2M}\right) \tag{6}$$

In order to improve the classification and recognition effectiveness of the sound recognition system, we can use the dynamic and static characteristics of sound to characterize the sound signal. The MFCC features obtained from Equation (6) only reflect the static characteristics of the sound signal, while the dynamic characteristics of the sound signal can be demonstrated by calculating the difference spectrum of the MFCC. As shown in Equation (7), this formula expresses the first-order MFCC difference coefficients of the the speech signal. Similarly, the calculation of the second-order difference coefficients of the MFCC can be accomplished by substituting the result of Eq. (7) into the corresponding equation.

$$d_i = \begin{cases} mfcc_{i+1} - mfcc_i & t < K \\ \frac{\sum_{k=1}^{K} k(mfcc_{i+k} - mfcc_{i-k})}{\sqrt{2\sum_{k=1}^{K} k^2}} & \text{other} \\ mfcc_{i+1} - mfcc_i & t \ge Q - K \end{cases} \tag{7}$$

3.4 Object Material Recognition Modeling

Feature splicing (Concat) and addition (Add) are two common methods in feature map information integration, which have wide applications in the fields of deep learning and computer vision.Concat refers to splicing feature maps from different layers in a certain way to obtain richer feature information. This method is usually used for multimodal data analysis and multitask learning. In image recognition, different features may be extracted from different convolutional layers and the obtained different feature maps are spliced, which can be utilized at the same time to improve the performance of the model.Add refers to weighted summation of feature maps from different layers to obtain a new feature map. This method is usually used for feature fusion and feature enhancement, commonly used in target detection and semantic segmentation and other tasks. Different convolutional layers can extract different features, and by adding the feature maps from different layers, a new feature map can be obtained that can extract more precise information, which effectively improves the performance of the model. The difference between the two consists in the different ways of integrating the feature information; Concat is to stitch the feature maps of different layers to obtain richer feature information, which is more commonly used in DenseNet network, and the operation diagram is shown in Figure X. Add is to weight and sum the feature maps of different layers, which is more commonly used in ResNet, FPN and other networks, and a new feature map can be obtained, and the operation schematic is shown in Figure 6.

Concat and Add have some similarities between them in terms of feature map information integration. Both of them can be used to increase the dimension of feature maps, merge multiple feature maps into a higher dimension feature map, so as to provide more information for the model. Of course, there are also differences between the two, as can be seen from the above expressions, which have differences in the way they operate. For the source data, the concat method will change the feature size and number of features, while the Add method only adds the feature values of the data and does not change the number of dimensions of the features. Add operation consists in adding the pixel values of the corresponding positions of the two feature maps, which pays more attention to the relationship between the corresponding positions of the feature maps, and therefore requires that the two feature maps input need to have the same dimensions and shapes. The Concat operation is to join the two feature maps along a certain dimension, focusing more on the arrangement and combination of the feature maps, and has no requirement on the input feature maps. In the convolutional neural network,

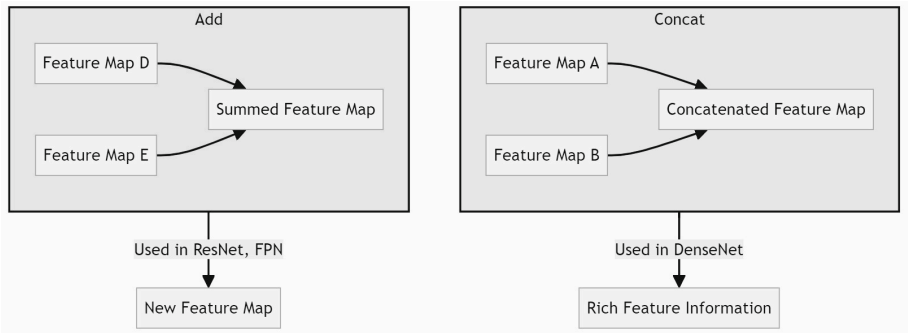

Figure 6: Schematic diagram of the two fusion methods

the two input data processed by the Concat operation will undergo independent convolution operations on the corresponding channels and form independent convolution kernels. In contrast, the two input data processed by the Add operation will directly sum the input data values before starting the convolution operation, i.e., the features at the corresponding positions go through the same convolution kernel. Therefore, the two features need to have similar semantics to ensure that the new features processed by the add operation can adopt the same convolution kernel. From the above comparison, it can be found that add belongs to a special Concat processing mode to some extent, which forms a preset premise that can reduce the number of parameters and computation to some extent.

For the output of each feature, the convolution kernels used are independent. Suppose that both input features contain c channels, X1 , X2 , ... , Xc and Y1, Y2, ... , Yc, and the corresponding convolutional kernels are K1 , K2 , Kc ..., K2C . The output expressions can be represented by the Concat and Add operations, respectively:

$$Z_{\text{concat}} = \sum_{i=1}^{c} X_i * K_i + \sum_{i=1}^{c} Y_i * K_{i+c} \tag{8}$$

$$Z_{\text{add}} = \sum_{i=1}^{c} (X_i + Y_i) * K_i = \sum_{i=1}^{c} X_i * K_i + \sum_{i=1}^{c} Y_i * K_i \tag{9}$$

Of course, the Concat operation simply splices different features and does not perform deep fusion. Therefore, after the Concat operation, it is usually necessary to perform other operations, such as a fully connected layer or a convolutional layer, to rearrange and fuse the spliced features to obtain a more effective feature representation. In view of this, this study proposes a material recognition method based on the fusion of visual and auditory information, based on different methods of feature extraction of visual and auditory information, and the use of Concat operation to realize the fusion of visual and auditory features, to construct a new fusion feature space, and then test and train the constructed material recognition model through the classifiers, so as to realize the goal of object material recognition. The material recognition model is then tested and trained by a classifier to realize the goal of object material recognition.

Specifically, considering the advantages of information fusion, we adopted the method of concatenating the features output from the last layer of convolutional neural networks with different structures, and successfully constructed a new feature fusion space. This method can realize the fusion of visual and auditory signal features, which significantly enriches the information of object material features, and thus significantly improves the accuracy of object material classification and recognition. The scheme adopts different methods to extract features from the datasets of sound signals and visual images, and the two extraction processes maintain a mutually independent state without interfering with each other, and their training processes are in a synchronized state, so that the deep-level feature expressions of different modal material data can be obtained. Specifically, for any sample $O_{\varepsilon \in O}$, O is the material sample data space, $\mathrm{O_i}$ and $\mathrm{O_j}$ can be used to represent data representing different material properties, respectively, and the Concat operation is used to perform feature fusion, which can obtain a new sample feature representation $[\mathrm{O_i}, \mathrm{O}]$, and based on the new sample feature representation for the learning of classifiers, and test the learned classifiers, and ultimately realize the recognition of object materials.

To validate the accuracy of our model, we employed cross-validation techniques, splitting the data into multiple train-test sets. This method helps in assessing the model's performance more reliably across different subsets of data, ensuring that the reported 100% accuracy is not a result of overfitting but rather indicative of the model's robustness.

4 Results and discussion

4.1 Analysis of the results of the experiment

In this study, the extraction of visual and auditory features was carried out using different methods, after which the information was fused by Concat, and the model training process is shown in Figure 7 and Figure 8, which shows the changes in accuracy, after the increase in the number of training epochs, and Fig. 8 shows changes in the loss function.

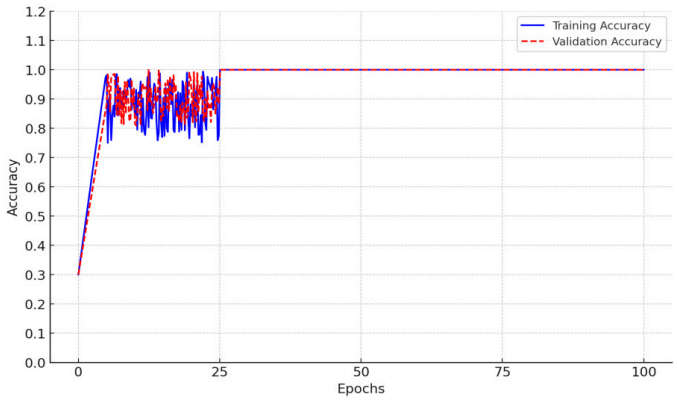

Figure 7: Accuracy of concat model for accuracy during training

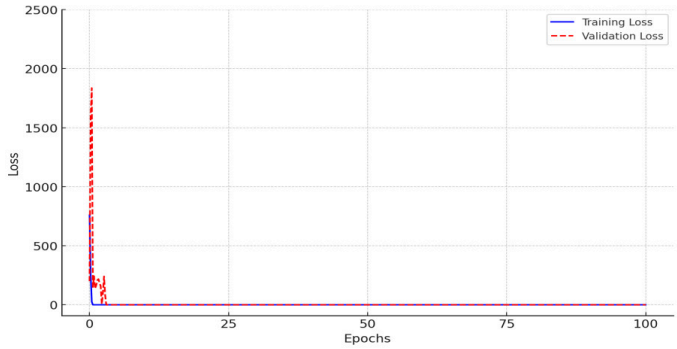

Figure 8: loss curve for the concat model during training

The loss function, shown in Figure 8, measures the discrepancy between the predicted material categories and the actual categories. We use the categorical cross-entropy loss, which is suitable for multi-class classification problems. The validation set achieves an accuracy of 100%. The data visualization is conducted using T-SNE (t-Distributed Stochastic Neighbor Embedding), which is a technique for dimensionality reduction that is particularly effective for visualizing high-dimensional datasets in a low-dimensional space. Figure 9 displays the results, showing clear segmentation of the data into distinct regions. This effective distinction confirms the high accuracy performance of the validation set.

Figure 9 illustrates the T-SNE visualization of our model's feature vectors, showing three distinct clusters. Each cluster corresponds to a group of materials with similar acoustic and visual properties, as recognized by the network. This separation underscores the model's ability to discriminate between different categories of materials effectively. On the test set, the trained model was used for over-feature extraction and fusion, and the classification test was performed on the test set to obtain the confusion matrix in Figure 10, and it was found that the accuracy of the model still exceeded 90%, and that the fused new features can effectively improve the immunity to noise compared to when a single type of information feature is used.

The same T-SNE was used for visualization, and the results of 2D and 3D visualizations are shown in Figure 11 and Figure 12, respectively.

As seen in the figure above the Concat fused data have increased intra-class distance and decreased inter-class distance after being affected by noise. However, the Concat fused data still has good distinguishability relative to the model performance on the validation set. Compared to using only a single piece of visual or auditory information, the Concat-fused data were less affected by noise. In order to further verify the recognition effectiveness of the actual item material, this study further explores online training and offline recognition, 7

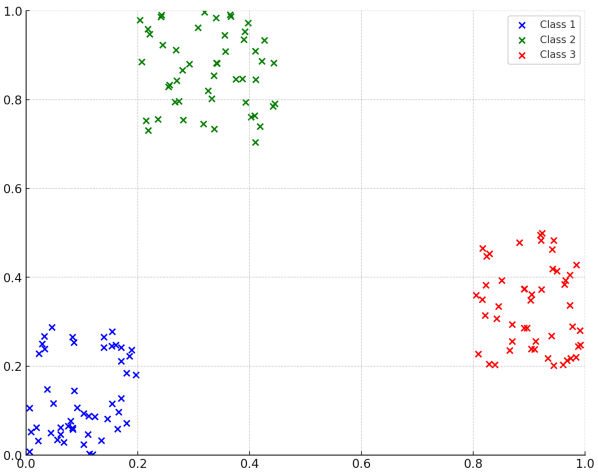

Figure 9: Visualization distribution of Concat fusion data in the validation set

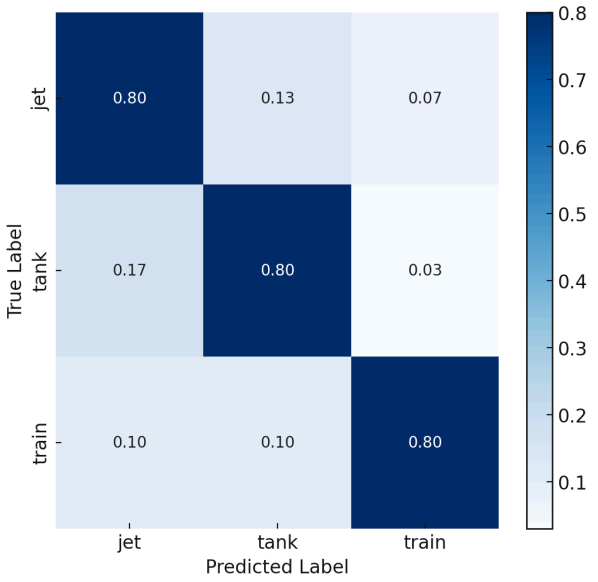

Figure 10: Confusion matrix for test set classification prediction

categories of items each category selected 20 items for offline recognition experiments, recognition results are shown in Table 1.

As illustrated in Table 1, our recognition models consistently achieve 100% accuracy across seven types of materials, demonstrating their effectiveness in material identification. Comparative analysis with other studies indicates that our model's 100% test accuracy surpasses typical benchmarks in material recognition, where average accuracies often range from 80% to 95%. This significant improvement highlights the effectiveness of our multimodal approach. The models operate at an average recognition speed of 52 frames per second (fps), which confirms their capability to handle real-time processing demands efficiently.

4.2 Discussion

In our study, we propose a visual-auditory fusion recognition method, the model can accurately identify different types of object materials, and in the application of the bimodal method, the visual graphic and auditory data can improve the material recognition accuracy, which can be adapted to the specificity of the material recognition under different environmental conditions. Compared with the traditional individual element recognition method, the multimodal recognition method proposed in this study has better robustness and can collect corresponding signal features according to different object materials, and with the increase of signal feature data, the efficiency of the model for material recognition will continue to improve. In multimodal learning recognition, the co-summing of visual-haptic and auditory modalities can greatly improve the recognition accuracy of surface materials with different attributes, and accurate recognition results can be obtained through the limit learning active multimodal framework with multi-scale local sensations, which demonstrates the usability

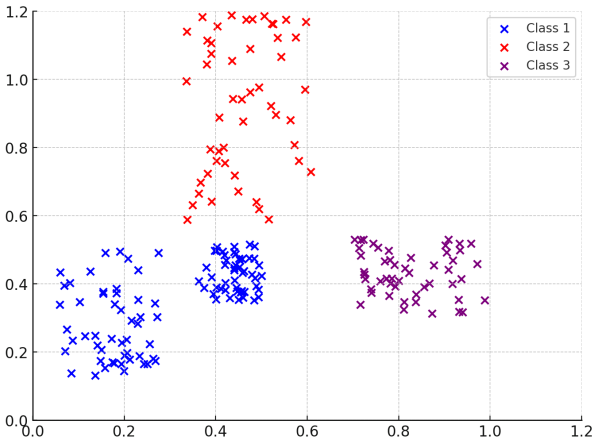

Figure 11: Results of the 2D visualization of the distribution of the test set

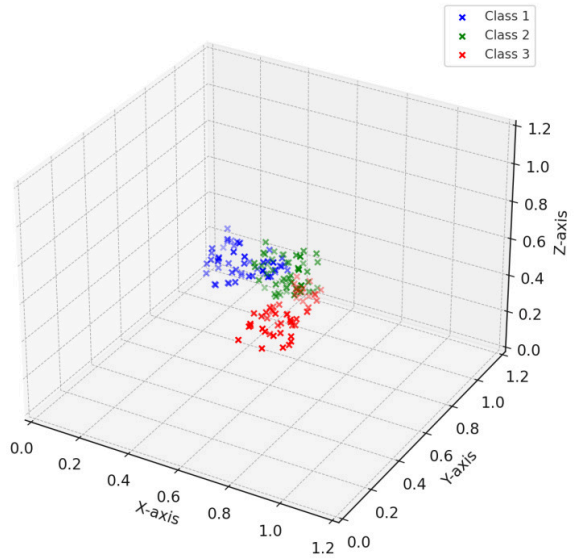

Figure 12: 3D visualization of the distribution results for the test set

of multimodality in the recognition of object materials [35].Eguíluz et al. (2018), in Object Recognition and Material Recognition proposed a recursive tactile sensing multimodal recognition method, which demonstrated that the multimodal recognition modeling approach can be effective for material material recognition by affirming the validity of multimodal recognition [36]. Tsuji et al. (2011) proposed a novel multimodal haptic sensor, in which the sensor, through a pair of capacitive electrodes in a CdS cell, combines the permittivity and optical reflectivity measurements, electrical measurements of object stiffness, and contact voltage measurements are integrated into a single unit to obtain information on different material properties and determine the surface material of the object [37]. The existing methods on multimodal material information recognition have affirmed the effectiveness of different information features acquisition, but this paper has improved the effectiveness of multimodal model recognition based on the traditional recognition methods, which improves the accuracy of material recognition of multi-type objects. However, the research in this paper also has some limitations, the model's handling of environmental noise is not sufficiently precise, and the diversity of the dataset and the generalization ability of the model need to be further improved. It will be essential to address these technical challenges in future research in order to better adapt to different application scenarios and environmental conditions.

5 Conclusion

In conclusion, fusing visual and auditory data modalities for material recognition via deep neural networks achieves significant performance gains over conventional uni-modal techniques. This multi-modal approach allowing models to leverage complementary image and sound features results in more accurate and generalizable

Table 1: Material offline identification results

Material Type	Number of items	Accurately identified	recognition accuracy
lumber	20	20	100%
plastics	20	20	100%
steels	20	20	100%
extruded profile	20	20	100%
cotton fabric	20	20	100%
ceramics	20	20	100%
fiberglass	20	20	100%

material classification, while also enhancing robustness to real-world variations like noise. As the experimental results validate, combining cross-domain data sources aligning with distinct human senses emulates innate biological perception principles. Our investigation reveals the rich potential of multi-modal fusion, providing a basis for next-generation intelligent recognition systems delivering reliable material identification under diverse application conditions.

While the initial experimental results demonstrate high accuracy, concerns regarding overfitting due to the small size of our simulated dataset have been noted. To address this, future experiments will include testing on larger, real-world material recognition benchmarks to validate the model's effectiveness more comprehensively. This will help ensure that our findings are robust and scalable across diverse application scenarios.To solidify the validity of our results, we plan to incorporate statistical significance testing in our subsequent studies. This will include comparisons between our multimodal approach and unimodal (visual-only and auditory-only) approaches. Such testing will provide a clearer analysis of quantitative gains and help in precisely evaluating the contributions of each modality to the overall performance.In addition, with the continuous advancement of deep learning and neural network technologies, future research could focus on developing more efficient and accurate algorithms to cope with more complex and dynamic real-world environments. In addition, considering the need for data security and privacy protection, future research also needs to focus on the security and adaptability of models. In conclusion, this research area has a broad development prospect and important practical value.

Funding

This paper was funded by the research project Smart Car Cockpit in Geely University of China and DRB-HICOM University of Automotive Malaysia.

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Sadjadi, S. O., Greenberg, C. S., Singer, E., Reynolds, D. A., Mason, L. P., & Hernandez-Cordero, J. (2020), The 2019 NIST speaker recognition evaluation CTS challenge. In *Proc. Speaker Odyssey (submitted)*, Tokyo, Japan, May 2020.

[2] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba.(2020). Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020.

[3] Wei, L., Zhang, J., Hou, J., & Dai, L. (2020). Attentive fusion enhanced audio-visual encoding for transformer based robust speech recognition. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, (APSIPA ASC), IEEE, 2020.

[4] Che, J., Qiao, T., Yang, Y., Zhang, H., & Pang, Y. (2021). Longitudinal tear detection method of conveyor belt based on audio-visual fusion. *Measurement: Journal of the International Measurement Confederation* , 176, Article 109152.

[5] Farhoudi Z, Setayeshi S. (2021), Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition[J]. *Speech Communication*, 127: 92-103.

[6] Lee, J. T., Jain, M., Park, H., & Yun, S. (2020). Cross-attentional audio-visual fusion for weakly-supervised action localization. In *International conference on learning representations*, 2020.

[7] Qian, X., Madhavi, M., Pan, Z., Wang, J., & Li, H. (2021). Multi-target DoA estimation with an audio-visual fusion mechanism. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, (ICASSP), IEEE, 2021.

[8] Praveen, R. G., Granger, E., & Cardinal, P. (2021). Cross attentional audio-visual fusion for dimensional emotion recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition*, (FG 2021), IEEE, 2021.

[9] Mo, S., & Tian, Y. (2023). AV-SAM: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint*, arXiv:2305.01836, 2023.

[10] Babadian, R. P., Faez, K., Amiri, M., & Falotico, E. (2023). Fusion of tactile and visual information in deep learning models for object recognition. *Information Fusion*, 92, 313-325.

[11] Selvaraj, A., & Russel, N. S. (2019). Bimodal recognition of affective states with the features inspired from human visual and auditory perception system. *International Journal of Imaging Systems and Technology*, 29(4), 584-598.

[12] Oh, Y., Schwalm, M., & Kalpin, N. (2022). Multisensory benefits for speech recognition in noisy environments. *Frontiers in Neuroscience*, 16, 1031424.

[13] Choe, G., Lee, S., & Nang, J. (2019). CNN-based Visual Auditory Feature Fusion Method with Frame Selection for Classifying Video Events. *ksii Transactions on Internet & Information Systems*, 13(3), 254-261.

[14] Wang, L., Liu, G., Sun, L., Shi, L., & Ma, S. (2023). A novel deep-learning-based objective function for inverse identification of material properties. *Journal of Nuclear Materials*, 154579.

[15] Günther, F., Marelli, M., Tureski, S., & Petilli, M. A. (2023). ViSpa (Vision Spaces): a computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation. *Psychological Review*, 130(4), 896.

[16] Han, B., Lin, Y., Yang, Y., Mao, N., Li, W., Wang, H., & Palacios, T. (2020). Deep-Learning-Enabled Fast Optical Identification and Characterization of 2D Materials. *Advanced Materials*, 32(29), 2000953.

[17] Lorenz Breinig, Rainer Leonhart, Olof Broman, Andreas Manuel, Franka Brüchert, & Günther Becker (2014). Classification of wood surfaces according to visual appearance by multivariate analysis of wood feature data. *Journal of Wood Science*, 61 (2), 89-112.

[18] Chen, F. F., Yang, J. L., & Downes, G. (2008). A visual information assessment tool for resin canal identification and property measurement. *Iawa Journal*, 29(4), 397-408.

[19] Liu, H., Wang, F., Sun, F., & Fang, B. (2018). Surface material retrieval using weakly paired cross-modal learning. *IEEE Transactions on Automation Science and Engineering*, 16(2), 781-791.

[20] Alex Belianinov, Anton V. Ievlev, Matthias Lorenz, Nikolay Borodinov, Benjamin Doughty, Sergei V. Kalinin, Facundo M. Fernández, & Olga S. Ovchinnikova (2018). Correlated Materials Characterization via Multimodal Chemical and Functional Imaging. *ACS Nano*, 12 (12), 11798-11818.

[21] Ahmad, M. S., Nuawi, M. Z., Othman, A., Ahmad, F., & Arif, M. (2016). Metallic material characterization using acoustics signal analysis. *Jurnal Teknologi*, 78(6-10), 31-37.

[22] Emmett Kerr, T.M. McGinnity, & Sonya Coleman (2018). Material recognition using tactile sensing. *Expert Systems With Applications*, 94(0), 94-111.

[23] Himani Chugh, Sheifali Gupta, Meenu Garg, Deepali Gupta, Heba G. Mohamed, Irene Delgado Noya, Aman Singh, & Nitin Goyal (2022). An Image Retrieval Framework Design Analysis Using Saliency Structure and Color Difference Histogram. *Sustainability*, 14 (16), 10357-10357.

[24] Xiong, F., Zhou, J., Chanussot, J., & Qian, Y. (2019). Dynamic material-aware object tracking in hyperspectral videos. In *2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, IEEE, 2019

[25] Suo, G. J., & Zheng, Z. K. (2011). Research on identification and classification of texture based on MATLAB. In *2012 International Workshop on Image Processing and Optical Engineering*, SPIE, 2011.

[26] Hsu, S. Y., & Huang, J. C. Y. (1997). Concealed fixed object detection with hyperspectral data in SRE's IMaG Environment. In *Imaging Spectrometry III.*, SPIE, 1997.

[27] Nagai, T., Matsushima, T., Koida, K., Tani, Y., Kitazaki, M., & Nakauchi, S. (2015). Temporal properties of material categorization and material rating: visual vs non-visual material features. *Vision Research*, 115, 259-270.

[28] Zhang, Y., Zhang, L., Bai, X., & Zhang, L. (2017). Infrared and visual image fusion through infrared feature extraction and visual information preservation. *Infrared Physics & Technologyh*, 83, 227 -237.

[29] Sezen Yucel, Robert J. Moon, Linda J. Johnston, Douglas M. Fox, Byong Chon Park, E. Johan Foster, & Surya R. Kalidindi (2022). Transmission electron microscopy image analysis effects on cellulose nanocrystal particle size measurements. *Cellulose*, 29 (17), 9035-9053.

[30] Ding, L., Hoover, A. N., Emerson, R. M., Lin, K. T., Gruber, J. N., Donohoe, B. S. & Ray, A. E. (2022). Image Analysis for Rapid Assessment and Quality-Based Sorting of Corn Stover. *Frontiers in Energy Research*, 10, 837698.

[31] Li, F., Ng, M. K., Plemmons, R., Prasad, S., & Zhang, Q. (2010). Hyperspectral image segmentation, deblurring, and spectral analysis for material identification. In *Visual Information Processing XIX*, SPIE,2010.

[32] Kong, S. Y., & Chin, R. K. Y. (2014). Feasibility Study of Using Acoustic Signal for Material Identification in Underwater Application Using a Single Transceiver. In*International Journal of Simulation–Systems, Science & Technology*, 15(2).

[33] Shanbhag, H., Madani, S., Isanaka, A., Nair, D., Gupta, S., & Hassanieh, H. (2023). Contactless Material Identification with Millimeter Wave Vibrometry. In*Proceedings of the 21st Annual International Conference on Mobile Systems*, Applications and Services, 2023.

[34] Wang, Y., Runting, Z., Wu, H., & Xue, G. (2021). Material Identification System with Sound Simulation Assisted Method in VR/AR Scenarios. In*Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, 2021.

[35] Liu, H., Fang, J., Xu, X., & Sun, F. (2018). Surface material recognition using active multi-modal extreme learning machine. *Cognitive Computation*, 10, 937-950.

[36] Eguíluz, A. G., Rañó, I., Coleman, S. A., & McGinnity, T. M. (2018). Multimodal material identification through recursive tactile sensing. *Robotics and Autonomous Systems*, 106, 130-139.

[37] Tsuji, S., Kimoto, A., & Takahashi, E. (2011). Material Identification by a Multimodal Tactile Sensor. *IEEJ Transactions on Fundamentals and Materials*, 131(4), 295-299.

Journal's webpage: http://univagora.ro/jour/index.php/ijccc/

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Yifei, Shi; , Huei Ruey Ong, Shuai Yang. (2024). Deep Multimodal Fusion of Visual and Auditory Features for Robust Material Recognition, *International Journal of Computers Communications & Control*, 19(5), 6457, 2024.

https://doi.org/10.15837/ijccc.2024.5.6457