IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 3, 2022, accepted May 9, 2022, date of publication May 13, 2022, date of current version May 26, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3174874

Object Description Using Visual and Tactile Data

PENG ZHANG[1], MAOHUI ZHOU[2], DONGRI SHAN[2], ZHENXUE CHEN[3], AND XIAOFANG WANG[1]

[1]School of Electrical Engineering and Automation, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
[2]School of Mechanical Engineering, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
[3]School of Control Science and Engineering, Shandong University, Jinan 250061, China

Corresponding author: Peng Zhang (zp@qlu.edu.cn)

This work was supported in part by the Project of Shandong Provincial Major Scientific and Technological Innovation under Grant 2019JZZY010444 and Grant 2019TSLH0315, in part by the Project of 20 Policies to Facilitate Scientific Research at Jinan Colleges under Grant 2019GXRC063, and in part by the Natural Science Foundation of Shandong Province of China under Grant ZR2020MF138.

ABSTRACT With the development of vision and haptic sensor technologies, robots have become increasingly capable of perceiving their external environment. Although machine vision and haptics have surpassed humans in some aspects of perception, it is difficult for robots to describe objects from multiple viewpoints using a combination of visual and haptic modalities. In this study, we use convolutional neural networks to separately extract visual and haptic features and then fuse these two types of features. Then, multitask learning is combined with multilabel classification to form a multitask-multilabel classification method. The developed method is used to identify the color, shape, material attributes, and class of an object from the visual-haptic fused feature vector. To verify the effectiveness of the proposed object description method, experiments are conducted on the PHAC-2 dataset and the collected VHAC dataset. The experimental results show that the proposed method produces the most accurate object descriptions with the smallest number of parameters.

INDEX TERMS Object description, machine vision, machine haptics, multimodal fusion, multitasking-multilabel.

I. INTRODUCTION

It is a common human behavior to perceive objects in visual and tactile ways and describe them verbally. The core of this behavior combines visual and tactile perception to make judgments about the appearances, materials, and categories of objects. The ability of computers to recognize objects from images surpassed that of humans [1] as early as 2015. Machine haptics also perform well in terms of object texture recognition [2]–[4] and material classification [5]–[7] tasks. However, vision-based appearance recognition and haptic-based material recognition are still separate research directions in the field of robotics, which results in robots being unable to form descriptions of objects through visual and haptic perception, as performed by humans.

Currently, there is no standardized dataset for training deep neural networks to provide visual and haptic-based object descriptions. To address this problem, we recreated the labels of the Penn Haptic Adjective Corpus 2 (PHAC-2) dataset [8] to describe objects in terms of their categories, colors, material attributes, and shapes. However, the haptic collection process of the PHAC-2 dataset is too complicated. Humans usually generate object descriptions by looking at and grasping objects. To explore whether robots can generate object descriptions by grasping and glancing, we collected the VHAC dataset and applied labels regarding four aspects: category, color, shape, and material attributes.

In this paper, we propose an object description method based on multimodal perception with mu ltitask-multilabel classification (MMM). As shown in Figure 1, in the multimodal feature extraction part, visual and tactile features are extracted using DenseNet169 [9] models and a 1D convolutional neural network, respectively, and then visual feature vectors and haptic feature vectors are concatenated into feature fusion vectors. In the feature classification part, a multitask-multilabel classification method is proposed. The classification method is a multilabel classification technique containing four subtasks that describe objects in terms of four aspects: category, shape, color, and material attributes. In this paper, the accuracy of an object description is measured using exact matching. Experimental results on the PHAC-2 and VHAC datasets show that the MMM method yields the

The associate editor coordinating the review of this manuscript and approving it for publication was Laura Celentano.

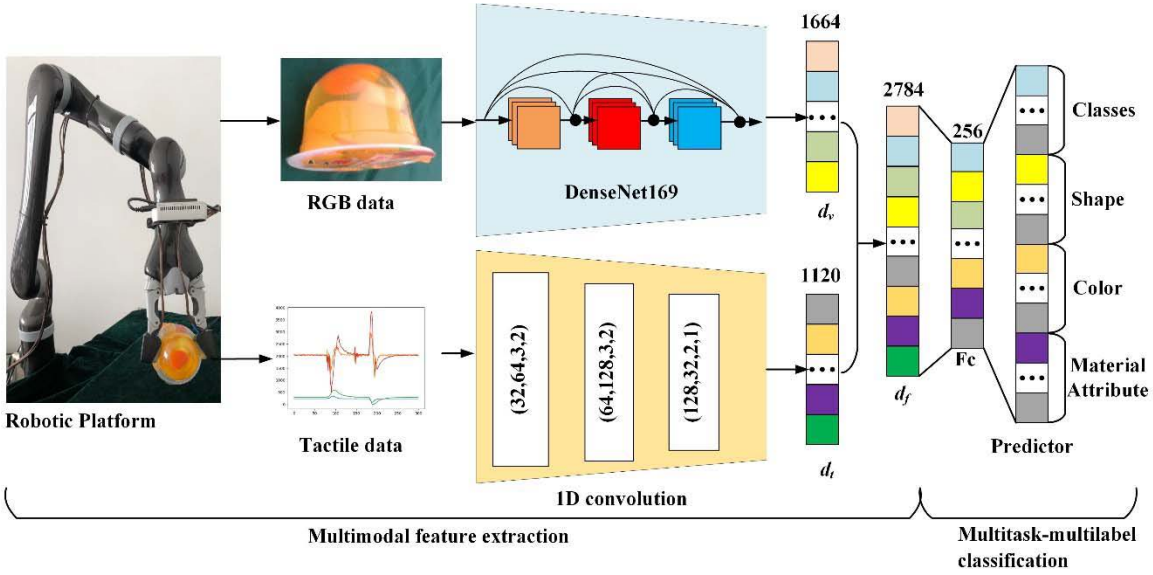

FIGURE 1. Object description model based on visual and tactile perception.

best object description results with the smallest number of parameters in comparison with other methods.

Our main contributions are as follows.

1) An object description method based on visual and tactile perception is proposed. Objects can be described in terms of four aspects: class, color, shape, and material attributes.
2) Experiments prove that the multitask-multilabel classification method can effectively balance each task and achieve the highest matching accuracy with respect to object description.
3) An object description dataset based on robot grasping is established, and the feasibility of forming object descriptions based on visual and tactile grasping is proven.

II. RELATED WORK AND BACKGROUND

Robots often use machine vision to recognize the appearances of objects in terms of their colors, shapes, sizes, and textures. In material classification [10], [11] and texture recognition [12], researchers have tried to build relevant image datasets to explore more application scenarios for machine vision. However, the images in these datasets do not show sufficient surface details for material classification and texture recognition.

Tactile sensors can obtain richer material information than vision through direct contact with objects, and thus, haptics are widely used in object texture recognition [2]–[4] and material classification [5]–[7] tasks. To improve the haptic cognitive abilities of robots, Chu *et al.* [8] collected haptic data using a BioTac haptic sensor and built the PHAC-2 dataset. In this dataset, 24 haptic adjectives were used to describe the material attributes of an object, which were identified by using a multicore learning classifier. Subsequently, zero-shot learning approaches [13], hierarchical extreme learning machines [14], and unsupervised learning methods [15] have been used for material property identification. These methods treat each material attribute as a binary classification problem in the classification process. Another line of research focuses on the correlations between different material attributes and treats material property classification as a multilabel classification problem [16], [17]. However, the disadvantage of tactile sensors is that the tactile contacts are localized and cannot perceive information about the appearance of an object.

Machine vision and machine haptics have advantages and shortcomings in object appearance recognition and material recognition, respectively. Therefore, multimodal deep learning methods based on visual-haptic fusion are used for object classification [18] and material property recognition [7], [19]. Experiments have shown that multimodal approaches are better than unimodal visual or haptic approaches. However, the above types of methods achieve object classification or material property recognition in a single-task learning manner. When faced with different objects composed of the same material and the same objects made up of different materials, single-task methods do not recognize and describe these objects from multiple perspectives.

Multitask learning can describe objects from multiple perspectives, but no related work has implemented an object description method based on visual and tactile perception using multitask learning methods. Current multitask classification methods have their own private networks and loss functions for each task. During the training process, there are competition problems between different tasks and different learning speeds. Therefore, a large amount of research has focused on multitask parameter sharing methods to reduce the competition between different tasks [20]–[22]. Studies

have focused on balancing the weights between different loss functions [23], but these methods attempt to achieve the best prediction for each task, ignoring how to make each task simultaneously output correct predictions.

III. MODEL INTRODUCTION

A. MULTIMODAL FEATURE EXTRACTION

Three main types of data fusion methods are available—early (data-level), intermediate (feature-level), and late (decision-level) approaches [24]. In [25], the author compared the three types of fusion methods on the PHAC dataset, and the experimental results showed that the intermediate fusion method produced better experimental results. In this paper, we use the intermediate fusion method to realize the fusion of tactile and visual data. The deep learning network model is designed in this paper as follows.

1) HAPTIC MODEL

The haptic data in the PHAC-2 and VHAC datasets consist of 1D time series. The input of the haptic model is a $C \times T$ matrix (see Section 4.1 for details), where C is the data dimensionality of the haptic data and T denotes the length of the haptic data.

Although Long Short Term Memory (LSTM) models [26] are generally quite skilled at addressing time series data, relevant comparative experiments have shown that LSTM models perform considerably worse than 1D convolutional neural networks [7]. As shown in figure 1, this paper uses a three-layer 1D convolutional neural network to extract tactile features. The parameters in the parentheses indicate the number of input channels, the number of output channels, the size of the convolution kernel, and the sliding step. A rectified linear unit (ReLU) function is employed as the activation function in each layer of the neural network.

2) VISUAL MODEL

The visual data in the PHAC and VHAC datasets are RGB images of objects taken from different angles. Visual and tactile sensations are complementary in terms of object description. Therefore, we need to find a visual model that works well with the haptic model. In this paper, DenseNet169 models are used as vision models, and three-layer 1D convolutional neural networks are used as haptic models for experiments. The experimental results of the DenseNet169 model are found to be better than those of the other models in terms of accuracy and stability (see Section 5.1 for details).

Because the numbers of images in both the PHAC-2 dataset and VHAC dataset are small, we use the MINC-2500 [11] dataset to pretrain the visual model. This dataset is a visual material recognition dataset containing 23 classes of objects with 2500 samples in each class.

3) VISUAL-HAPTIC FUSION

As shown in Figure 1, the last layers of the visual and tactile convolutional neural networks are evaluated separately to obtain a visual feature vector d_v and a haptic feature vector d_t. As shown in Equation (1), the number of features is concatenated to obtain a visual-haptic fusion feature vector d_f as follows:

$$d_f = [d_v^T + d_t^T]^T \tag{1}$$

Both the visual feature vector d_v and the tactile vector d_t are column vectors, and the sizes of the vectors are shown in Figure 1. In the model used in reference [7], the size of d_t was approximately four times that of d_v. The main task of this paper is to describe objects in terms of their visual and tactile aspects. Therefore, we adjust the network parameters of the tactile model to reduce the gap between d_t and d_v.

B. MULTITASK-MULTILABEL CLASSIFICATION METHOD

Previous work treated each material property as a binary classification problem [7], [8], [14], [15]. The 24 tactile adjectives of the PHAC-2 dataset require 24 training and test sets. This method of binary classification learning for each adjective is not flexible enough, the overall parameters are large, and the correlation between the different tactile adjectives is fragmented. Recently, some scholars used a multi-label classification method to classify tactile adjectives when studying the correlation between tactile adjectives, but their research was limited to the tactile aspect [16], [17]. Inspired by their research, this paper introduces a multi-label classification method into object description based on a multimodal model. Due to the stronger scalability of multi-label classification, we add shape, color and object category labels to PHAC-2.

In this paper, objects are described in terms of four aspects: color, shape, material attributes, and category. A multitask-multilabel classification method is proposed to coordinate the individual tasks and simultaneously produce accurate description results. The method combines four tasks to form a multilabel classification approach. The specific implementation is shown in Equation (2), where the color label y^1, the shape label y^2, the material attribute label y^3 and the category label y^4 of the object are combined to form a multilabel classification vector with four subtasks as follows:

$$y^m = [y^1 + y^2 + y^3 + y^4] \tag{2}$$

The multitask-multilabel classification method is a multilabel classification task containing four subtasks, and this paper defines the multilabel classification loss L_{Ml} as follows.

$$L_{Ml}(x, y) = -\frac{1}{C} * \sum_i y[i] * \log\left((1 + \exp(-x[i]))^{-1}\right) + (1 - y[i]) * \log\left(\frac{\exp(-x[i])}{(1 + \exp(-x[i]))}\right) \tag{3}$$

where x indicates the model output, y is the supervision label, $x[i]$ represents the value of x, and $y[i]$ represents the value of y. Here, $y[i] \in \{0, 1\}$, i $\in \{0, \cdots, x_n - 1\}$, and $x_n - 1$ is the number of output elements.

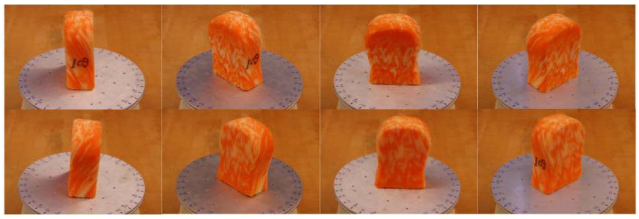

FIGURE 2. Visual data of the orange sponge in the PHAC-2 dataset.

From Equation (3), it can be derived that during the back-propagation process of the neural network, when $y[i] = 1$, $x[i]$ is a relatively large positive number to ensure that $LMl(x, Y)$, takes the minimum value; Moreover, when $y[i] = 0$, $x[i]$ is a relatively large negative number to ensure that $L_{Ml}(x, y)$ takes the minimum value. Therefore, the multil-abel classification output needs to be adjusted for positive and negative samples by choosing a threshold value. As shown in Equation (4), this paper uses 0 as the threshold value for $x[i]$ as follows:

$$x[i] = \begin{cases} 1x_1[i] > 0 \\ 0x_1[i] < 0 \end{cases} \tag{4}$$

IV. EXPERIMENT

A. DATASET FORMATION

1) PHAC-2 DATASET

The PHAC-2 dataset was collected by Chu *et al.* [8] and contains the visual information and tactile signals of 53 types of objects. The image collection process was performed as follows: each object was placed on a rotating platform and photographed every 45 degrees of rotation to obtain 8 images from different angles.

Figure 2 shows the visual data of an object in the PHAC-2 dataset. A BioTac haptic sensor was installed at the end of the fingers of a PR2 robot to perform four processes, including squeezing, holding, slow sliding, and fast sliding, on each object. The collected haptic data included low-frequency fluid pressures (P_{DC}), high-frequency fluid vibrations (P_{AC}) core temperatures (T_{DC}) core temperature changes (T_{AC}) and 19 electrode impedances (E1 $\cdots$ E19). P_{AC} signals were collected 10 times for each object with a sampling frequency of 2200 Hz, and the rest of the signals were sampled at 100 Hz. Figure 3 shows the original data collected by the tactile sensor, where the shaded part is valid haptic.

A total of 24 haptic adjectives were used to describe the haptic sensations of the objects in the dataset, and the haptic adjectives for each object were determined using 36 experimenters. As all 24 adjectives in PHAC-2 dataset are material attribute adjectives of objects, the dataset cannot meet the requirements of this paper to describe objects from the two aspects of object color and shape. To address the visual description shortcomings of the PHAC-2 dataset, a total of 10 volunteers were employed to determine the colors and shapes of objects based on the visual color images of 53 objects in the dataset. As shown in Tables 1 and 2, a total of 19 colors and 6 shapes were used to describe the objects.

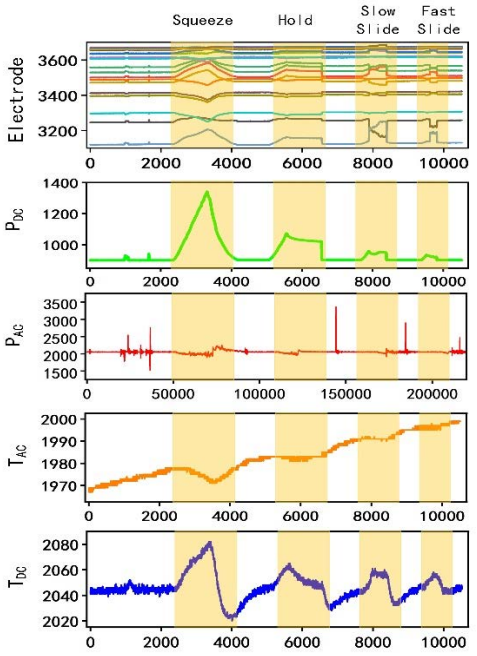

FIGURE 3. Tactile data of the orange sponge in the PHAc-2 dataset.

TABLE 1. Color labels of the PHAC-2 dataset.

Silver	Light blue	Brown	White	Black
Yellow	Dark blue	Multicolored	Deep red	Pink
Khaki	Gray	Off-white	Dark green	Transparent
Green	Red	Milky-white	Orange	

TABLE 2. Shape labels of the PHAC-2 dataset.

Long strip	Cylindrical	Flattened	Bottled	Chunky	Folding shape

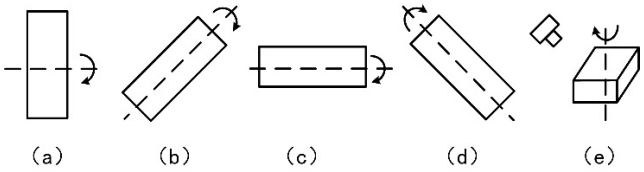

FIGURE 4. Visual data collection process.

2) VHAC DATASET

Humans recognize and perceive most of the properties of an object when they see and grasp it. To simulate the human grasping-based cognitive process, this paper collects the visual and haptic data of 22 grasped objects. The data collection platform is shown in Figure 1, with a RealSensor D435i camera mounted on the wrist of a Kinova robotic arm and a NumaTac haptic sensor mounted on the finger end of the arm.

During the visual data collection process, the wrist camera of the robotic arm first reaches above the object to shoot the object, as shown in Figure 4(a)-(d). The object is clockwise along the symmetry axis of the figure in each pose and rotates

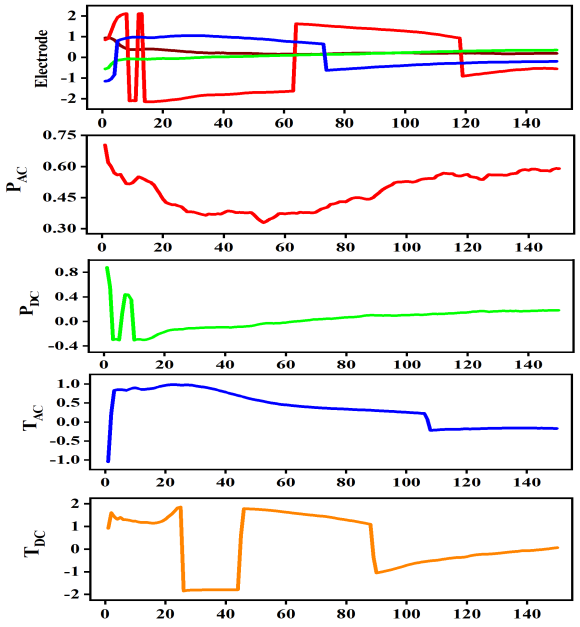

FIGURE 5. Tactile data after compression.

4 times at 90° each time. Then, the camera reaches obliquely above the object, as shown in Figure 4(e). During the shooting process, the object rotates 4 times clockwise along the symmetry axis in the figure, and each time, it rotates 90°. This method is used to simulate the situation of the robot arm approaching the object from different directions during the gripping process. Twenty images of each object with different angles are obtained, and 22 objects are shown in Figure 6.

During the haptic data collection process, the arm approaches each object vertically from above and performs the gripping action. The NumaTac tactile sensor collects low-frequency fluid pressures and high-frequency fluid vibration signals from the arm during the gripping process, with each object being gripped 20 times with a uniform texture and no substandial material changes. The P_{AC} sampling frequency is 2200 Hz, and the P_{DC} sampling frequency is 100 Hz. A grasp can be divided into three processes: clamping, holding and opening, where the clamping and opening times vary slightly between materials, and the holding time is 1.5 s. The clamping process is stopped when $P_{DC} = 270bit$. The conversion formula for the haptic sensor threshold and pressure is

$$P = (P_{DC} - offset) * 12.94\ Pa/bit \tag{5}$$

where off set $= 238\ Pa/bit$.

Because the NumaTac sensor does not contain temperature data or the 19 electrode impedances, the VHAC dataset does not contain diverse tactile adjectives, such as temperature and viscosity. As shown in Tables 3 and 4, the VHAC dataset contains 8 material attributes and 12 colors to describe an object. The VHAC dataset is the same as the PHAC-2 dataset in terms of its shape descriptions. The adjectival description

TABLE 3. Material attribute adjectives of the VHAC dataset.

Soft	Slightly hard	Hard	Compressible
Resilient	Smooth	Uneven	Textured

TABLE 4. The color labels of the VHAC dataset.

Transparent	With text	Black	Silver	Yellow	White
Red	Orange	Khaki	Multicolored	Pink	Green

of each object is determined for both datasets by using 10 experimenters.

B. DATA PREPROCESSING

1) TACTILE DATA

In this paper, we follow the tactile data processing method of Gao *et al.* [7]. Because PAC is sampled at a higher rate than other signals, we first downsample it to 100 Hz to match the sampling rate of other signals. Then, the useless parts in the original data are removed, and the tactile data of the four tactile operations are retained. The electrode impedances (E1$\cdots$ E19) are downsampled to 4-dimensional impedance signals. Finally, tactile signals are obtained from $P_{DC}, P_{AC}, T_{DC}T_{AC}$ and the 4-dimensional impedance signals. As shown in Equation (6), the data of each dimension are normalized to obtain S'. In the formula, $\bar{S}$ is the mean value of the data, and σ is the standard deviation of the data.

$$s' = \frac{S - \bar{S}}{\sigma} \tag{6}$$

Figure 5 shows the tactile data of the fast sliding process after data processing. The four tactile action processes are concatenated to obtain 32-dimensional tactile data. The size of the tactile data for each object is $C_P \times T_P$, where $C_P = 32$ denotes the number of tactile channels and $T_P = 150$ denotes the length of the data.

Unlike the PHAC-2 dataset, this paper treats a grasp as a haptic exploration action. The haptic data of the VHAC dataset contain only P_{DC} and P_{AC} First, we downsample P_{AC} to 100 Hz. Then, the tactile data obtained from each grasping process are sampled to a length of 300. Finally, the haptic data of the left and right hands are concatenated to obtain a haptic representation $C_V \times T_V$ where $C_V = 4$ and $T_V = 300$. Figure 7 shows the haptic data collected during each object grasping process, where the abscissa represents the data length and the ordinate represents the pressure. The larger amplitudes in the figure are the P_{AC} and P_{DC} of the two tactile fingers. According to the tactile data, it can be seen that the amplitude of a soft object is small (the signal rises and falls slowly), and the amplitude of a hard object is large (the signal changes quickly). Because the cleaning sponge is thinner than the other objects and it is quickly compacted during the clamping process, a higher amplitude and larger signal changes occur while grasping the cleaning sponge. In addition, it was found that the paper cup is hard at first,

FIGURE 6. Tactile data of the 22 categories of objects in the VHAC dataset.

and as the pressure increases, a certain level of deformation is caused by vibration. The vibration signal change process of a smooth object is flatter, while the rising and falling of the vibration signal of a rough object are accompanied by a small vibration.

2) VISUAL DATA

Due to the small numbers of visual images in the PHAC-2 and VHAC datasets, an image enhancement technique is used in this paper to make the samples of the training set as diverse as possible for the training process. The specific implementation steps are as follows. First, the brightness, saturation and contrast levels of the images are randomly adjusted to between 70% and 130% of their original values, and the image size is adjusted to 300×300 pixels. Then, each image is randomly cropped to 224×224 pixels. Finally, the images are randomly flipped horizontally with a 50% probability. During the test, each image is resized to 300×300 pixels, and a 224×224 pixel image is extracted from the center as the model input.

C. EXPERIMENTAL SETTINGS

1) FEATURE EXTRACTION

In this section, the superiority of the multimodal feature extraction model is analyzed in terms of three aspects.

First, this paper verifies whether adjusting the size of d_t and d_v can improve the accuracy of the model. In the model used in reference [7], the size of d_t is approximately four times that of d_v. We reproduce the model and set $d_v : d_t \approx 1.5 : 1$.

Then, the feature extraction models of different visual models are compared. GoogLeNet [27], ResNet152 [1], MnasNet [28] and DenseNet169 models are used as comparative visual feature extraction models.

Finally, related studies have compared four multimodal fusion methods, low-rank multimodal fusion (LFM), the mixture of experts (MoF) approach, late fusion (Late), and intermediate fusion (Mid), on the PHAC-2 dataset [26]. To validate the effectiveness of the visual-haptic fusion method proposed in this paper, the same comparison experiments as those conducted in [25] are performed on the PHAC-2 dataset.

2) FEATURE CLASSIFICATION

To verify the effectiveness of the proposed multitask-multilabel classification method, this paper replicates the current mainstream hard parameter sharing [20], soft parameter sharing [21], cross-stitch network [22] and multiple single-task joint classification methods. The current multitask classification methods have separate outputs and loss functions for each task. The main network architectures of these multitask classification methods are the same, but their differences are in the sharing mechanisms used among the various tasks.

In the multitask classification approach for classifying object colors, shapes and material properties, the loss functions used for all three tasks are multilabel loss functions (Equation (7)). The loss function for the object classification task L_c (classification) is defined as follows:

$$L_c(x, y) = -x[y] + \log\left(\sum_j \exp(x[j])\right) \tag{7}$$

where x denotes the model output, y denotes the category label, and x[j] represents the value of x.

The total loss function of the multitask classification method is defined as follows:

$$\begin{aligned} L_{Mt}(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4) &= L_{Ml}(x_1, y_1) \\ &+ L_{Ml}(x_2, y_2) + L_{Ml}(x_3, y_3) + L_C(x_4, y_4) \end{aligned} \tag{8}$$

where (x_1, y_1), (x_2, y_2), (x_3, y_3), and (x_4, y_4) are the output and supervised labels for color, shape, material attribute and classification tasks in the multitask classification method, respectively.

The multiple single-task joint classification approach separately trains multiple networks for different tasks and then combines the results of the individual network models. This approach has an independent visual-haptic fusion model and a single-task classification network for each task. This method uses L_{Ml} as the loss function for the color, shape, and

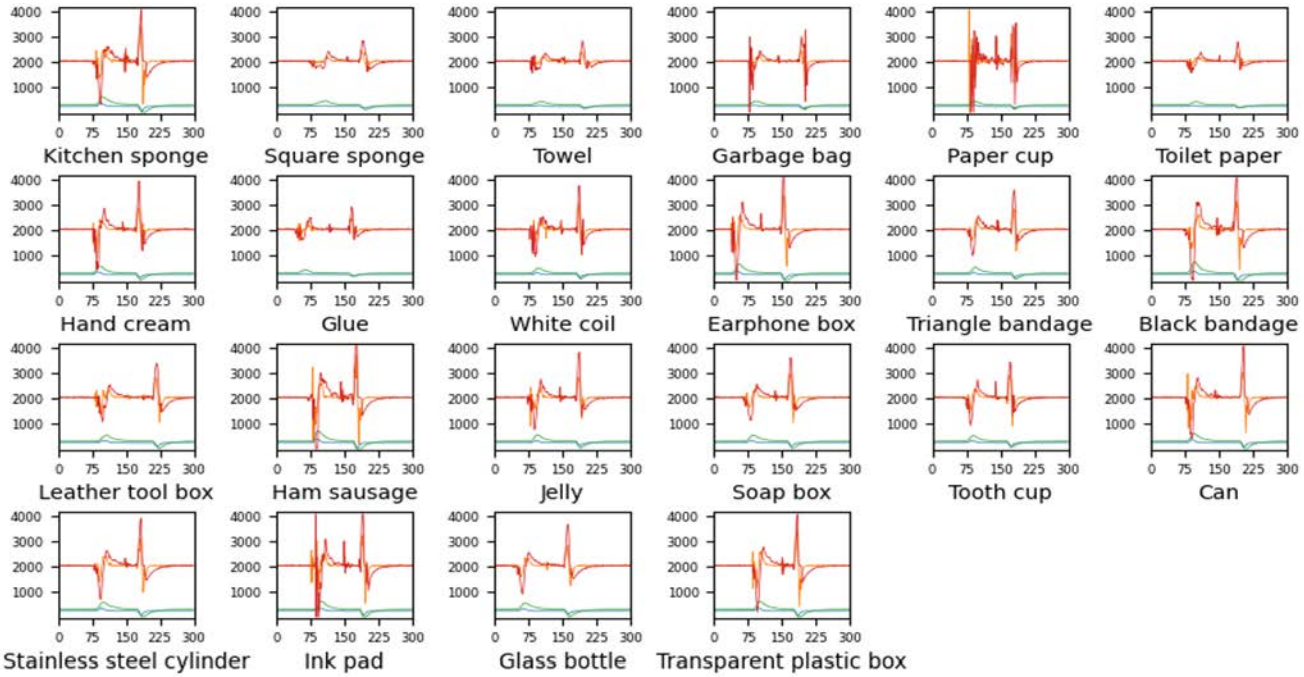

FIGURE 7. Tactile data of the 22 categories of objects in the VHAC datas.

material property classification tasks, and employs L_C as the loss function for the object category classification task.

The visual and haptic data are paired when input into the neural network during the experiments. To ensure a fair comparison experiment, the number of fully connected layer neurons and the activation function are the same in all classification methods. During the training process, 20% of the neurons of the fully connected layer are randomly deactivated to enhance the robustness of the network. To accelerate the training speed of the neural network, the output of the fully connected layer is normalized, which is shown in Equation (9) [29] as follows:

$$y = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}}\gamma + \beta \tag{9}$$

where $\gamma = 1, \beta = 0, \epsilon = 1e-5$, and γ and β are the parameter vectors of size C that can be learned (C is the input size).

The VHAC dataset contains 20 visual and haptic samples per object, which are divided into two training and test sets at a ratio of 1:9. Each object in the PHAC-2 dataset contains 8 images and 10 tactile data points. Thus the dataset has only eight valid visual and tactile data pairs. Each object keeps one visual and tactile dataset as a test set and the rest of the data are used as a training set. To ensure the accuracy of the experimental data, five different pairs of test and training sets are randomly formed.

In this paper, we use the PyTorch deep learning framework to build the neural network model, and the model is run on an Nvidia Tesla V100 graphics card during training and testing. Throughout the training process, an Adam-based parameter optimization method is used, where the batch size is set to 4, the learning rate is set to 0.00002, the remaining 10 parameters are set to their default values, and the number of epochs is set to 250.

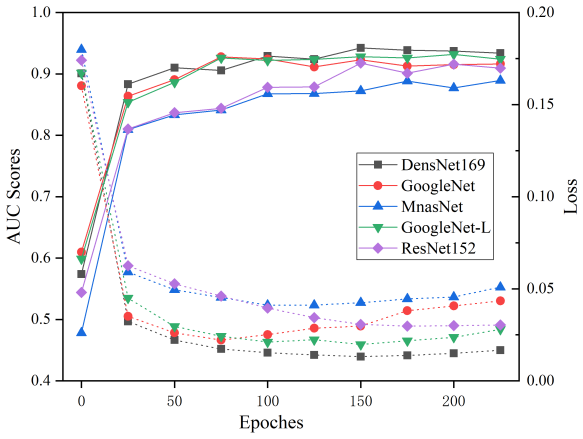

FIGURE 8. Comparison results produced by different visual models.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. FEATURE EXTRACTION RESULTS

Figure 8 shows the results of the material attribute classification experiments yielded by different visual models on the PHAC-2 dataset. A total of 225 epochs are used for training in the experiment, and one test is performed every 25 epochs. It can be seen from the figure that the DenseNet169 model has better accuracy and stability than the other models. It is

TABLE 5. Comparison of the visual-haptic fusion methods.

Method	Test	Test	Test	Test	Test	Average
LFM	0.896	0.898	0.902	0.928	0.900	0.901±0.012
MoF	0.923	0.919	0.919	0.923	0.927	0.922±0.003
Late	0.923	0.924	0.922	0.923	0.923	0.923±0.0006
Mid	0.929	0.922	0.922	0.927	0.925	0.925±0.0028
M	0.916	0.938	0.935	0.871	0.916	0.915±0.024
MS	**0.979**	**0.974**	**0.971**	**0.98**	**0.985**	**9.78**±0.0049

worth noting that in the figure, GoogLeNet is the visual-tactile fusion network model reproduced in reference [7], and GoogLeNet-L is the model with a reduced d_t size. It can be seen in the figure that the loss increases and the accuracy of the GoogLeNet model decreases during the later period, indicating network overfitting. The above comparison proves that the sizes of the visual and tactile feature vectors should not be too different in the visual-tactical fusion model.

As shown in Table 5, this paper compares the proposed method with other visual-haptic fusion methods on the PHAC-2 dataset. Due to the uneven distribution of the material attribute samples of objects in the PHAC-2 dataset, the area under the curve (AUC) score is used in this paper as the evaluation criterion for material attribute identification. The first five rows of Table 5 show the results and average values obtained in the five material attribute classification experiments by the LFM, MoF, Late, Mid and M methods on the PHAC-2 dataset. The visual-haptic fusion model proposed in this paper is titled ''multimodal & single-task'' (MS) in Table 5. The data comparison shows that the MS method performs much better than the other methods in terms of material attribute classification, which proves that the visual-haptic fusion method is better than the other methods for feature extraction.

B. FEATURE CLASSIFICATION RESULTS

The comparison of the proposed multitask-multilabel classification method with other multitask classification methods on the PHAC-2 and the VHAC datasets is shown in Figure 9. The images in the first and second rows are the experimental results achieved on the PHAC-2 and the VHAC datasets, respectively. The pictures in each row are the experimental material property, shape, color, category and matching accuracy results. A total of 250 epochs are used for training in the experiment, and one test is performed every 50 epochs.

In the Figure 9, the AUC scores is used as the evaluation criterion for the color, shape, and material attributes, and the rate of correctness is used for the categories. The data in the table are the averages calculated over five experiments. The task in this paper is to form accurate object descriptions, and it is necessary to make the four tasks as accurate as possible while simultaneously performing prediction. We use precision matching to measure the performance of the different methods on the object description task. In this paper, precision matching refers to the percentage of the four tasks that are simultaneously correctly predicted.

TABLE 6. Comparison of parameter quantities for different models.

Dataset	Model	Parameters
PHAC	Single-task union	52.96 M
	Hard parameter sharing	15.4 M
	Soft parameter sharing	15.4 M
	Cross-Stitch Networks	15.4 M
	Multitask-multilabel	**13.26 M**
VHAC	Single-task union	53.56 M
	Hard parameter sharing	16.03 M
	Soft parameter sharing	16.03 M
	Cross-Stitch Networks	16.03 M
	Multitask-multilabel	**13.40 M**

Experiments are performed with hard parameter sharing, soft parameter sharing and cross-stitch networks in the fully connected classification part to achieve parameter sharing. As shown in Table 6, the cross-stitch networks and other operations have little effect on the parameters of the overall network, and after conducting an analysis, it is found that the numbers of parameters of the three methods only differ by approximately 0.001M. Since the tactile data in the VHAC dataset are longer than those in the PHAC-2 dataset, the overall number of parameters is larger for each method when training on the VHAC dataset.

C. DISCUSSION

From the multiple sets of comparison data in Figure 9, it can be seen that the proposed multitask-multilabel classification method is slightly better than the other methods in terms of individual colors, shapes, and material attributes but significantly better than the other methods in terms of precision matching. The higher matching precision indicates that the multitask-multilabel classification method can better simultaneously coordinate the given tasks. In this paper, objects are described in terms of four aspects, and once an error occurs with respect to one aspect, it leads to inaccurate descriptions, especially when the descriptions of the color, shape and material attributes of an object correspond incorrectly to the categories of the object. Such incorrect descriptions do not have any meaning.

Each multitask-multilabel task is part of the multilabel classification procedure, and different tasks use the same loss function. The network automatically learns the interrelationships between different tasks during the training process to obtain accurate object description results. On the other hand, the supervised labels of different tasks are merged, and the merged labels increase the label differences between different objects. For example, for different objects with the same color and shape, performing the visual and color classification tasks without merging the labels tends to produce two identical

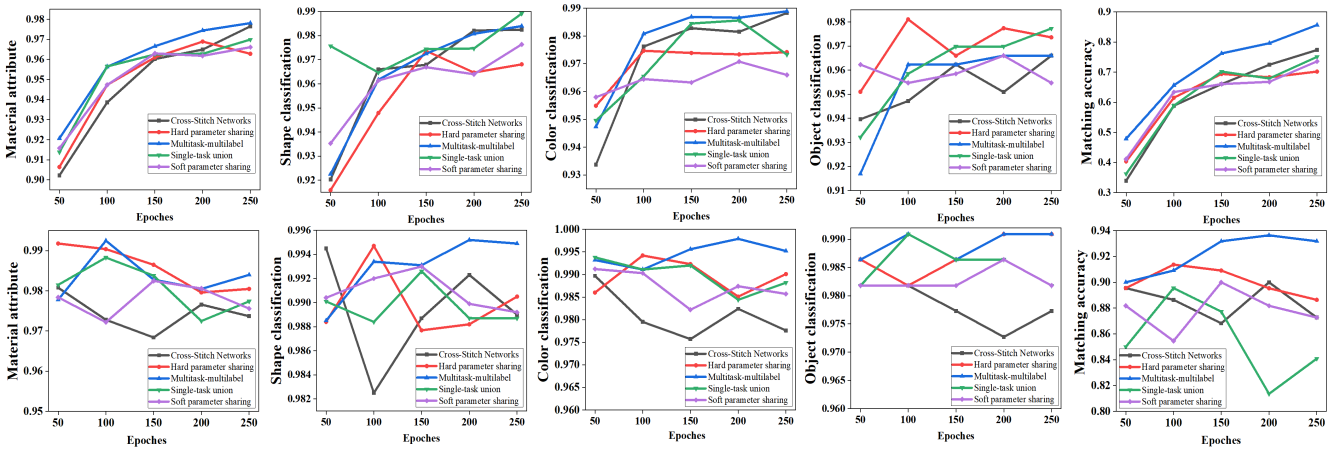

FIGURE 9. Comparison of multitask classification methods. The images in the first and second rows are the experimental results obtained on the PHAC-2 and VHAC datasets, respectively.

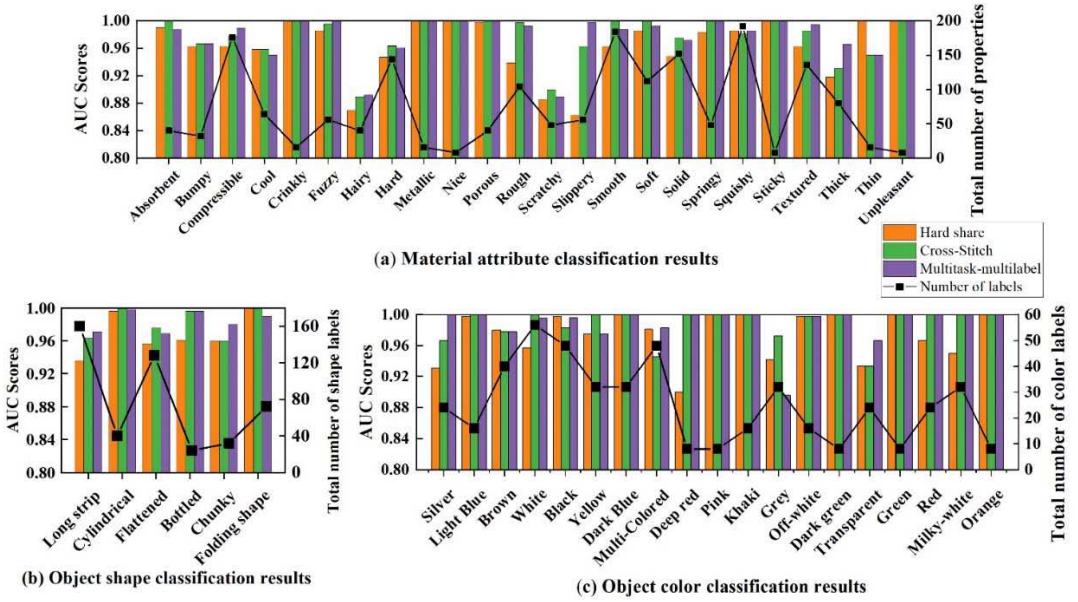

FIGURE 10. Classification results of the three methods on the 1:7 partition of the PHAC-2 dataset.

objects, which affects the accuracy of the classification task. After merging the labels into one label, the overall label differences make the neural network parameter optimization goal more explicit.

To verify the possibility of forming object descriptions after the robot sees and grasps an object, experiments are conducted on the VHAC dataset, and the results are shown in the second line of Figure 9. Due to the relative simplicity of the VHAC dataset, the five models achieve high classification accuracy by the 50th epoch. However, as training progresses, it can be found that the multitask-multilabel method has stable performance and has huge advantages in terms of accurate matching.

Compared with the four fast object actions of squeezing, holding, slowly sliding, and quickly sliding in the PHAC-2 dataset, the process of grasping objects in the VHAC dataset is more common in practical applications. From the experimental results, this process can be achieved by using visual observation and simple tactile grasping operations when facing simple object description tasks in daily life.

Figure 10 shows the classification results of the hard parameter sharing, cross-stitch network, and multitask-multilabel approaches for each task on the PHAC-2 dataset. From the material attribute classifications produced by the three methods, all three methods perform poorly regarding the classification of the "hairy" and "prickly" attributes. This may be because the distinction between these two material attributes or between them and other attributes is not obvious. The hard parameter sharing method does not perform well in terms of "thick", "smooth" and "slippery" attribute

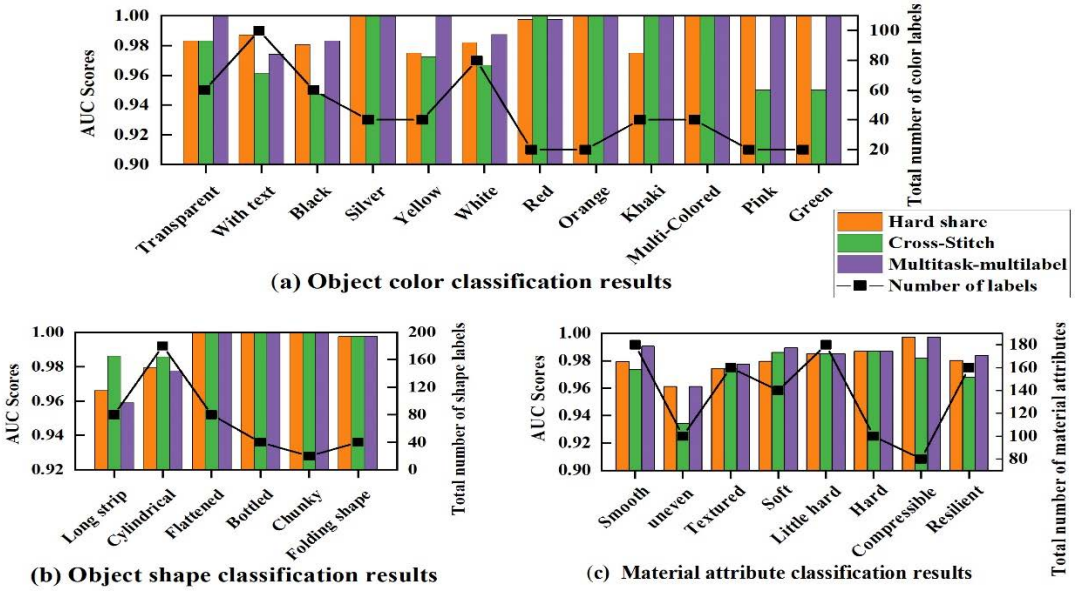

FIGURE 11. Classification results obtained by the three methods on the 1:9 partition of the VHAC dataset.

classification but is significantly better than the other methods regarding ''thin'' attribute classification. The hard parameter sharing method is worse than the other two methods at shape and color classification. The multitask-multilabel method is worse than the other methods at classifying ''gray'' but significantly better than the other methods at classifying ''silver'' and ''transparent''. Overall, the multitask-multilabel method performs best, and the hard parameter sharing method performs worst.

In addition, a commonality can be found from the line graph in Figure 10, where the number of labels is not related to the classification accuracies of the three tasks. For example, ''precise'', ''sticky'', ''unpleasant'', ''bottle'', ''pink'', ''dark red'', and ''orange'' are all classified in each task. ''Pink'', ''dark red'', and ''orange'' have the fewest numbers of labels for each task, but all of them have nearly 100% correct classification rates. Although the numbers of these labels are small, the features of these labels are more distinguishable from other features, and the network model can easily learn these features and distinguish them from other features.

Figure 11 shows the classification results obtained for each task on the VHAC dataset by the hard parameter sharing, cross-stitch network, and multitask-multilabel approaches. From the figure, it can be seen that the cross-stitch networks perform worst in terms of color and material property classification and best in terms of shape classification. The comparison shows that the number of labels has a large impact on the cross-stitch networks. The hard parameter sharing method and the multitask-multilabel method perform better overall.

The combined experimental results obtained on the two datasets show that the multitask-multilabel approach performs best. The hard parameter sharing method performs the worst on the PHAC-2 dataset. The cross-stitch networks perform worst on the VHAC dataset. The comparative experimental evidence produced by the two datasets shows that keeping the labels in parallel between different tasks when performing the multitask-multilabel classification method can reduce the competition between different tasks and achieve the best classification results for each label in a balanced way.

VI. CONCLUSION

This paper proposes an object description method for robot arms. The method includes two parts: feature extraction and feature classification. In the feature extraction part, a previously developed visual-haptic fusion method is improved and compared with other visual-haptic fusion methods. The experiments prove that the improved visual-haptic fusion method in this paper is better than other methods. In the feature classification part, a multitask-multilabel classification method is proposed and compared with other multitask classification methods. It is demonstrated that the multitask-multilabel method can achieve the most accurate classification results with the smallest number of parameters while utilizing the same feature extraction network. Finally, this paper measures the accuracy of object description by using the precision matching effect. Experiments on the modified PHAC-2 and VHAC datasets show that the MMM method can better coordinate tasks and achieve 5%-10% higher accuracy for the matching precision than the other methods. Experiments on the VHAC dataset prove the feasibility of robot grasping and describing objects based on vision and touch.

This paper explores the field of robot object description. Similar to the field of image captioning, keywords such as objects, actions and scenes in images are extracted

and formed into description sentences. This paper hopes that robots can use vision and touch to explore objects autonomously. Then the object is described from four aspects: color, shape, material properties, and class of the object. In future work, we hope to build a larger visual and tactile dataset to enable robots to automatically generate diverse description sentences.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[2] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, and R. Fuentes, "ViTac: Feature sharing between vision and tactile sensing for cloth texture recognition," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 2722–2727, doi: 10.1109/ICRA.2018.8460494.

[3] F. Wang, H. Liu, F. Sun, and H. Pan, "Fabric recognition using zero-shot learning," *Tsinghua Sci. Technol.*, vol. 24, no. 6, pp. 645–653, Dec. 2019.

[4] G. Cao, Y. Zhou, D. Bollegala, and S. Luo, "Spatio-temporal attention model for tactile texture recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 9896–9902, doi: 10.1109/IROS45743.2020.9341333.

[5] Z. Erickson, N. Luskey, S. Chernova, and C. C. Kemp, "Classification of household materials via spectroscopy," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 700–707, Apr. 2019.

[6] M. Strese, L. Brudermueller, J. Kirsch, and E. Steinbach, "Haptic material analysis and classification inspired by human exploratory procedures," *IEEE Trans. Haptics*, vol. 13, no. 2, pp. 404–424, Apr. 2020, doi: 10.1109/TOH.2019.2952118.

[7] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, "Deep learning for tactile understanding from visual and haptic data," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 536–543, doi: 10.1109/ICRA.2016.7487176.

[8] V. Chu, I. McMahon, L. Riano, C. G. McDonald, Q. He, J. M. Perez-Tejada, M. Arrigo, T. Darrell, and K. J. Kuchenbecker, "Robotic learning of haptic adjectives through physical interaction," *Robot. Auton. Syst.*, vol. 63, pp. 279–292, Jan. 2015.

[9] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[10] L. Sharan, R. Rosenholtz, and E. H. Adelson, "Accuracy and speed of material categorization in real-world images," *J. Vis.*, vol. 14, no. 9, p. 12, Aug. 2014.

[11] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3479–3487, doi: 10.1109/CVPR.2015.7298970.

[12] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3606–3613, doi: 10.1109/CVPR.2014.461.

[13] Z. Abderrahmane, G. Ganesh, A. Crosnier, and A. Cherubini, "Haptic zero-shot learning: Recognition of objects never touched before," *Robot. Auto. Syst.*, vol. 105, pp. 11–25, Jul. 2018.

[14] F. Li, H. Liu, X. Xu, and F. Sun, "Haptic recognition using hierarchical extreme learning machine with local-receptive-field," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 3, pp. 541–547, Mar. 2019.

[15] B. A. Richardson and K. J. Kuchenbecker, "Improving haptic adjective recognition with unsupervised feature learning," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 3804–3810, doi: 10.1109/ICRA.2019.8793544.

[16] H. Liu, F. Sun, D. Guo, B. Fang, and Z. Peng, "Structured output-associated dictionary learning for haptic understanding," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 47, no. 7, pp. 1564–1574, Jul. 2017, doi: 10.1109/TSMC.2016.2635141.

[17] H. Wu, X. Liu, S. Fang, Z. Yi, and X. Wu, "Leveraging multi-label correlation for tactile adjective recognition," in *Proc. 3rd Int. Conf. Robot., Control Autom. Eng. (RCAE)*, Nov. 2020, pp. 122–126, doi: 10.1109/RCAE51546.2020.9294444.

[18] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual–tactile fusion for object recognition," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 2, pp. 996–1008, Apr. 2017, doi: 10.1109/TASE.2016.2549552.

[19] Z. Abderrahmane, G. Ganesh, A. Crosnier, and A. Cherubini, "Visuo-tactile recognition of daily-life objects never seen or touched before," in *Proc. 15th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Nov. 2018, pp. 1765–1770, doi: 10.1109/ICARCV.2018.8581230.

[20] E. Eaton and M. Lane, "Modeling transfer relationships between learning tasks for improved inductive transfer," in *Machine Learning and Knowledge Discovery in Databases*, vol. 5211, no. 1. Berlin, Germany: Springer, 2008, pp. 317–332.

[21] L. Duong, T. Cohn, S. Bird, and P. Cook, "Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 845–850.

[22] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3994–4003, doi: 10.1109/CVPR.2016.433.

[23] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7482–7491, doi: 10.1109/CVPR.2018.00781.

[24] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[25] M. Bednarek, P. Kicki, and K. Walas, "On robustness of multimodal fusion—Robotics perspective," *Electronics*, vol. 9, no. 7, p. 1152, Jul. 2020.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[28] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "MnasNet: Platform-aware neural architecture search for mobile," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2815–2823, doi: 10.1109/CVPR.2019.00293.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn(ICML)*, 2015, pp. 448–456.

PENG ZHANG received the B.S. degree in automatic and the M.S. degree from the Shandong Institute of Light Industry, Jinan, China, in 2003 and 2000, respectively, and the Ph.D. degree in detection technology and automation from Tongji University, Shanghai, China, in 2010. He is currently an Associate Professor with the School of Electrical Engineering and Automation, Qilu University of Technology (Shandong Academy of Sciences), Jinan. His current research interests include intelligent detection and control technology, robotics, and numerical control technology.

MAOHUI ZHOU received the B.S. degree from the Qilu University of Technology (Shandong Academy of Sciences), Jinan, China, in 2019, where he is currently pursuing the master's degree. His research interests include machine vision, machine haptics, and multimodal fusion.

DONGRI SHAN received the M.S. degree in mechanical manufacturing and automation from Shandong Industrial University, Jinan, in 2000, and the Ph.D. degree in mechanical engineering from Zhejiang University, Hangzhou, in 2003. He is currently a Professor with the School of Mechanical Engineering, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China. His current research interests include intelligent manufacturing and advanced numerical control equipment.

ZHENXUE CHEN received the B.S. degree in automatic from the School of Electrical Engineering and Automation, Shandong Institute of Light Industry, Jinan, China, in 2000, the M.S. degree in computer science from the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Image Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, in 2007. From 2012 to 2013, he was a Visiting Scholar with Michigan State University, East Lansing, Michigan, USA. He is currently a Professor with the School of Control Science and Engineering, Shandong University. He has published over 100 papers in refereed international leading journals/conferences, such as IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, INFORMATION SCIENCES, NEUROCOMPUTING, NEURAL COMPUTING AND APPLICATIONS, and *SP-IC*. His research interests include image processing, pattern recognition, and computer vision, with applications to face recognition.

XIAOFANG WANG received the M.S. degree in engineering signal and information processing from the Ocean University of China, Qingdao, in 2005, and the Ph.D. degree in electronic and information from Xi'an Jiaotong University, Xi'an, in 2018. She is currently a Lecturer with the School of Electrical Engineering and Automation, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China. Her current research interests include image processing, pattern recognition, and deep learning.