IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 20, 2019, accepted December 5, 2019, date of publication December 12, 2019,
date of current version December 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2959011

A Novel Multi-Modal One-Shot Learning Method for Texture Recognition

PENGWEN XIONG[1,2], KONGFEI HE[2], AIGUO SONG[1], (Senior Member, IEEE), AND PETER X. LIU[3], (Fellow, IEEE)

[1]School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China
[2]School of Information Engineering, Nanchang University, Nanchang 330000, China
[3]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada

Corresponding author: Pengwen Xiong (steven.xpw@ncu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61663027, Grant 61903175, and Grant 61773051, in part by the Jiangxi Natural Science Foundation under Grant 20181BAB211019 and Grant 20171ACB20007, in part by the China Postdoctoral Science Foundation under Grant 2018M642137, in part by the Jiangsu Province Postdoctoral Research Funding Plan under Grant 2018K024A, and in part by the Jiangxi Province Innovation Project for Graduate under Grant YC2019-S011.

ABSTRACT Most machine learning algorithms require a large set of training samples in order to achieve satisfactory performance. However, this requirement may be difficult to satisfy in practice. Take the one-shot learning (OSL) problem on texture recognition for example; the machine learning algorithm is difficult to achieve satisfactory results. In order to solve this problem, a novel multi-modal one-shot learning method for texture recognition is presented. First, in order to improve the robustness of identification and the anti-interference to noise, we addressed the nontravel texture recognition challenges of learn information about object categories from only one training sample by fusing varied modalities data, including image, sound and acceleration, which provides rich information regarding textures. Second, a novel dictionary learning model is designed, which contains the various modalities information, and can simultaneously learn the latent common sparse code for the different modalities. Third, an original regularization term is developed to enhance the degree of distinction of different classes. Furthermore, the common features of the three modalities are evaluated in the case of one-shot learning and used as the basis for feature selection. In the end, experiments were performed based on a data set which was published openly to validate the effectiveness of the presented method.

INDEX TERMS Texture recognition, dictionary learning, one-shot learning, multi-modal fusion.

I. INTRODUCTION

Texture recognition is indispensable in our daily life because texture is a basic property of object in the realistic world. With the rapid development of the computer vision, texture recognition problem can be solved well by neural network [1]–[4], but most of them dependent on huge data sets and large computing resources. However, there are often very few training samples available for neural networks. Under the condition of a few samples, samples of each category for training set are often very few, and sometimes it may be just only one sample, which has hindered the application of CNN. Therefore, the OSL problem of texture recognition at present is still a huge challenge.

OSL aims to learn information about object categories from a few, or even one, training samples. With the continuous development of CNN and its improved neural network, scholars begin to find that it relies too much on huge data sets. However, humans can often learn from fewer samples to make correct decisions, which is the original motivation of OSL. This problem was first proposed by Fei-Fei Le in [5], which utilizes a generative object category model and the traditional variational Bayesian framework for learning and representation of visual object labels from a small number of training samples. After several years of development, the OSL problem has been partially solved, especially the proposal of Siamese Network (SN). With the explosive development of neural networks in recent years, most methods proposed recently are based on neural networks. Combined with the method proposed in [5], [29] introduced a Hierarchical

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

Bayesian model, which based on composition and causality. Aiming at the problem of few-shot learning, a novel relation network, which uses a special constructed function to get all relationships, is proposed in [13] to address the problem of few-shot learning. Reference [31] developed a matching network, which can map a small set of annotations and an unannotated test sample to its corresponding label, can avoid the need of adjusting the new label category in the fitting process. Reference [32] designed a prototypical networks, which can learn a "good" mapping through neural network, project each sample into the same space, and then extract their mean as prototype for each type of samples. Reference [33] adopted the proposed GNN to solve OSL problems. Reference [34] utilized optimization as a model for few-shot learning, and achieved expected result. So that the OSL problem of face recognition and other important topics have been well solved, but it has some limitations [6]. For example, it is still difficult to obtain good accuracy under the limit condition of providing only one sample. Most of methods are based on prior knowledge and need to use a model learned from a complete data set. Moreover, most recent proposed methods are based on omniglot data set without considering the benefits brought by multi-mode fusion and rely too much on prior knowledge.

In addition, the problem of the lack of samples, to some extent, can be alleviated by the fusion of multi-modal measurements. The earliest instance of multi-modal fusion is audio-vision speech recognition (ASVR) [7], which was originally inspired by the McGurk effect [8]. Although ASVR cannot significantly improve the recognition rate in most cases, current studies have shown that it has superior effects when sound data contains noise signals [9], [10]. The challenge of multi-modal fusion can be summarized into five taxonomy [11]: 1) Representation; 2) Translation; 3) Alignment; 4) Fusion; 5) Co-learning. At present, some of the challenges can be solved perfectly. For example, the first challenge of representation can be solved by adopting different measures for different modes. For vision, a one-dimensional feature can be extracted by CNN, and audio features can be extracted by MFCC or I-vector. For data fusion, neural networks [12], support vector machine (SVM) [13], Maximum entropy model [14], Dempster-Shafer theory [15] and Bayesian inference [16] are commonly used. Moreover, the level of fusion can be divided in to three categories [10]: 1) Feature level or early fusion; 2) Decision level or late fusion; 3) Hybrid. Early fusion. Early fusion commonly adopted by researchers in the early stage of multi-modal fusion research, which can effectively fuse low-level features, so as to extract the features of common connections between diverse modalities. Hybrid approach is adapted by some researchers to fuse multi-modal data at the early level along with late level.

Sparse dictionary learning is another instrument of the multi-modal measurement fusion [17]. Sparse dictionary learning is a representation learning method which focus on finding a sparse representation of the input information. Each column vector in the dictionary is called atoms. Up to now, many methods of sparse dictionary learning are proposed, include MOD, K-SVD, LASSO and so on. Mairal *et al.* proposed a novel online optimization algorithm, based on stochastic approximations, which scales up gracefully to huge data sets with millions of input training examples [18]. In recent years, some researchers utilize sparse dictionary learning to conduct multi-modal data fusion of feature level, some researchers have done a lot of work on this. In [19], Liu *et al.* developed a visual-haptic fusion framework for object recognition. [20] solved the zero-shot learning (ZSL) problem of classifying untouched haptic object with the help of visual modality, the proposed method is based on the new intermodal regularization term. In [21], Liu *et al.*investigated the multi-modal data fusion method, and studied the sample-to-sample pairing relationship between sound and acceleration measurements. These methods are based on sparse dictionary learning.

In recent years, there has been some progress in the fusion of haptic and visual data. The field has a wide range of applications about robotics. Some dexterous hands (e.g. Barrett Hand) be equipped with tactile sensors, which facilitate the acquisition of tactile data. Another application of tactile sensors is texture recognition, accelerometer or pressure transducer measurements can be used to represent haptic information for texture recognition [22]–[24]. Recently, Strese *et al.* [25] introduced a data set which contains 69 textures,[1] the recording procedure of this data set can be divided into two steps: first, to use the camera preview to capture the surface image. Second, let the recording device hits the surface. After that, the operator can move the tool across the surface at will. As the database updates, the LMT haptic texture database contains some modalities measurements which include acceleration, friction force, image, metal detection, IR reflectance and sound. After all the offered modal measurements which is offered in LMT are evaluated, visual, accelerate and sound information show good performance. Fig. 1 respectively shows the multi-modal measurements of *G1FineAluminumMesh* and *G1RhombAluminumMesh,* include image, sound and acceleration. In particular, the observed value of acceleration is a three-dimensional data which contains x, y and z axes.

Image, sound and acceleration measurements are three classes of sensing modalities which frequently occur in texture classification. In this paper, the OSL problem for texture recognition is tackled by fusing the information of image, sound and acceleration, and the method of multi-mode fusion can avoid the dependence on prior knowledge. In this paper, a novel sparse model is design to fuse the data of different modalities, so as to form a new sparse coding vector with different modality information. The designed model can learn the common sparse coding to achieve the purpose of fusion, and it has got a wonderful result in our experiments. The contributions of this work can be summarized as follows:

[1] https://zeus.lkn.ei.tum.de/downloads/texture/

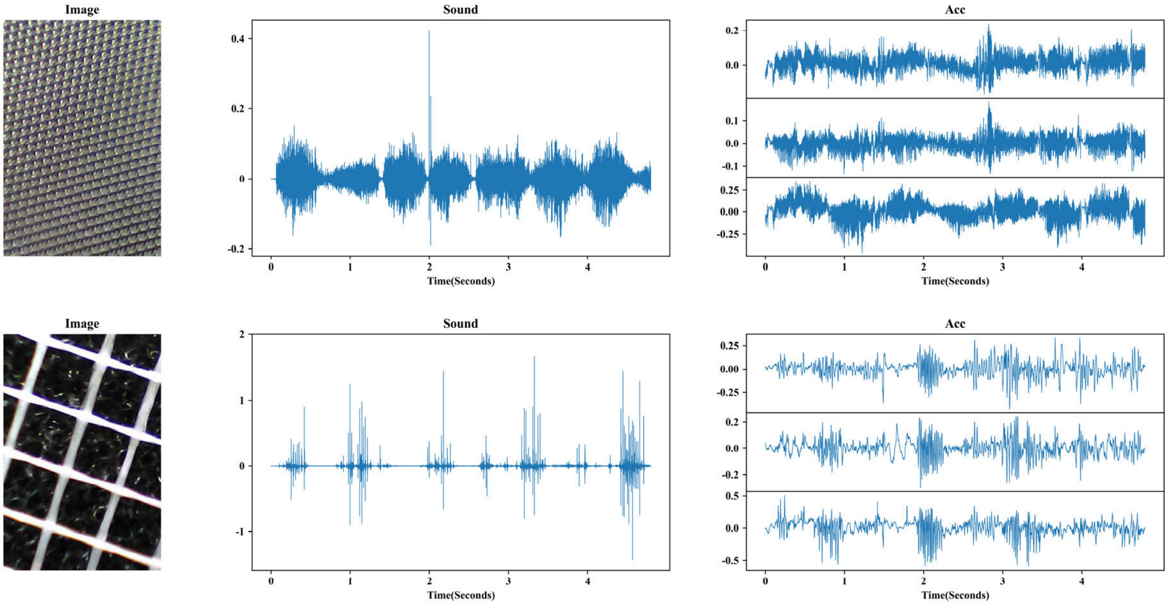


FIGURE 1. The image, sound and acceleration measurement of *G1FineAluminumMesh* and *G1RhombAluminumMesh*. It can be clearly observed that there are natural differences between different categories, different modalities of data. The samples are adopted from the LMT data set [25].

1) How to learn enough information from only one training sample, and systematically solve the problem of OSL about texture recognition.
2) A novel dictionary learning model is designed, which can utmost contain the various modalities information and simultaneously learn the latent common sparse code for the diverse modalities.
3) An original regularization term is developed to effectively promote the recognition effect for OSL problem.
4) The commonly used features of the three modalities are separately evaluated in the case of one-shot learning and the evaluation result is used as the basis for feature selection.
5) The experiment is performed based on a common data set to validation the effectiveness of our method, and the effects of various parameters on the results are also analyzed in detail.

The remainder of this work is presented as follows. The problem of multi-modal OSL is briefly described in Section II. In Section III, the novel dictionary learning model are introduced for OSL problem of the texture recognition. Section IV gives the result of evaluating the commonly used various features of vision, sound and acceleration modal. Section V introduces the classification experiments on our proposed method in different parameters. Conclusions are given in Section VI.

II. PROBLEM FORMULATION

The one-shot of texture recognition attempts to learn feature from the training set which comprise of only one or few samples from every class. The target is to acquire the true label of untrained samples. Our method, based on sparse dictionary learning, solves the problem of insufficient sample size by multi-modal fusion, and effectively avoids the accidental influence of single sample.

For convenience,

$$\begin{aligned} \boldsymbol{V} &= [\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N] \in \boldsymbol{R}^{d_V \times N}, \\ \boldsymbol{A} &= [a_0, \boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_N] \in \boldsymbol{R}^{d_A \times N} \end{aligned}$$

and

$$\boldsymbol{H} = [\boldsymbol{h}_0, \boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_N] \in \boldsymbol{R}^{d_H \times N}$$

are respectively employed to represent the feature vector of the training samples of image, sound and acceleration. The N presents the number of input training samples. The d_V, d_A and d_H respectively represent the feature dimensionality of input modality. It should be notified that the number of categories to be classified is equal to the number of input training samples. Each sample of any modality has their label from the number zero to N.

The mission of OSL can be formulated as predicting the true label, which lies in $\{0, 1, 2, \ldots, N\}$, of all input multi-modal measurements. The task of OSL holds the following three challenges:

1) The number of all input train samples of every class is only one. How to learn enough information from a single sample is a sea of troubles.
2) Each modality has unique feature, and the original data form is also different. Among them, image data are tensors with high-dimensional characteristics, sound signals are one-dimensional time series, and accelerometers are three-dimensional time series.

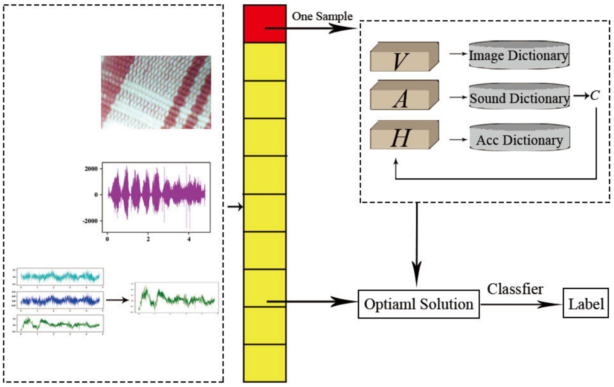


FIGURE 2. Illustration of the proposed dictionary learning method.

3) How to fuse multi-modal data to form a new feature, which must have high resolution for OSL.

In this paper, a new sparse dictionary learning model is specially designed to solve the above problems. A common sparse coding is tried to establish to eliminate the huge gap between the various modalities. To summary, the final task is that forming properly label of the input test instance $i \in \boldsymbol{R}^{d_V+d_A+d_H}$ by using the common sparse coding.

III. PROPOSED DICTIONARY LEARNING MODEL

A. PROPOSED MODEL

The sparse dictionary learning provides a perfect way to integrate the original features of different modalities to form a new sparse vector which contains original information. In OSL task, the sparse dictionary learning attempts to dispose the following optimization problem:

$$\min_{\boldsymbol{D_V},\boldsymbol{D_A},\boldsymbol{D_H},\boldsymbol{C}} \varsigma_r(\boldsymbol{D_V},\boldsymbol{D_A},\boldsymbol{D_H},\boldsymbol{C}) + \vartheta(\boldsymbol{C})$$
$$s.t.\, \boldsymbol{D_V} \in \boldsymbol{R}^{d_V\times K},\ \boldsymbol{D_A} \in \boldsymbol{R}^{d_A\times K},\ \boldsymbol{D_H} \in \boldsymbol{R}^{d_T\times K} \quad (1)$$

where $\boldsymbol{D}_V$, $\boldsymbol{D}_A$ and $\boldsymbol{D}_H$ are the dictionaries of diverse modalities, ς_r is the conventional reconstruction error term, ϑ is well-designed novel regularization term. As shown in Fig. 2, for different modalities, we try to make different modalities get the same sparse coding under their respective dictionaries. The $\boldsymbol{V}$, $\boldsymbol{A}$ and $\boldsymbol{H}$ can be coded as a spared vector which can be denoted as

$$\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3, \cdots, \boldsymbol{c}_N] \in \boldsymbol{R}^{K\times N}.$$

where K denotes the number of atoms in every dictionary, $\boldsymbol{C}$ denote the common sparse coding.

B. RECONSTRUCTION ERROR TERM

For sparse dictionary learning, the restoring ability, which transforms the spare coding vector to the original input data, of the over-complete base is largely reflected in the reconstruction error term. In this paper, the dictionary $\boldsymbol{D}_V$, $\boldsymbol{D}_A$ and $\boldsymbol{D}_H$ respectively are used to describe the input raw feature vectors of the single train sample of every class. The reconstruction error term can be formulated as

$$\varsigma_{\boldsymbol{r}} = \|\boldsymbol{V} - \boldsymbol{D}_V\boldsymbol{C}\|_F^2 + \|\boldsymbol{A} - \boldsymbol{D}_A\boldsymbol{C}\|_F^2 + \|\boldsymbol{H} - \boldsymbol{D}_H\boldsymbol{C}\|_F^2. \quad (2)$$

Obviously, only relying on the reconstruction error term is hard to solve the OSL problem well, so the carefully designed regularization terms needs to be utilized to deal with the investigated problem. The designed regular term contains two parts, the first part is conventional sparse regularization term, and another is inter-class regularization term. The conventional sparse regularization term makes the coding sparser, and another promote the recognition effect between different categories.

C. SPARSE REGULARIZATION TERM

1) CONVENTIOANL SPARSE REGULARIZATION TERM

The conventional sparse regularization term can be denoted as

$$\varsigma_s = \|\boldsymbol{C}\|_{1,1} \quad (3)$$

where the 1,1-norm $\|\ \|_{1,1}$ calculates the sum of one-norms of all columns, ς_s denotes the conventional sparse regularization term. Another common choice is zero-norm, but it is a NP-hard problem, and one-norm can achieve the effect of zero-norm under certain conditions.

2) INTER-CLASS REGULARIZATION TERM

The inter-class regularization term makes sparse coding of different label have better discrimination. A naive inter-class regularization term is

$$\varsigma_q = \frac{1}{\sum\limits_{i}^{|\boldsymbol{N}|}\sum\limits_{j=i+1}^{|\boldsymbol{N}|}\left\|\boldsymbol{c_i} - \boldsymbol{c_j}\right\|_2^2}. \quad (4)$$

The requirement is too strong in a way which it ignores the great contingency under the condition of single sample and difficult to calculate to some extent. After careful analysis, the inter-class regularization term can be considered as

$$\varsigma_z = \|\boldsymbol{C}\|_{0,2}, \quad (5)$$
$$\varsigma_o = \|\boldsymbol{C}\|_{1,2} \quad (6)$$

or

$$\varsigma_i = \|\boldsymbol{C}\|_\infty. \quad (7)$$

where the 0,2-norm $\|\ \|_{0,2}$ calculates the two-norm of the vector which is consisted by the zero-norms of all rows. 1,2-norm $\|\ \|_{1,2}$ calculates the two-norm of the vector which is consisted by the one-norms of all rows. But, the ς_z is excluded because it involves the zero-norm, which is NP-hard problem.

To sum up, the sparse regularization term is developed as

$$\vartheta(\boldsymbol{C}) = \lambda\varsigma_s + \delta\varsigma_\kappa \quad (8)$$

where ς_κ can be ς_o or ς_i, which represents the inter-class regularization term. Later sections will discuss their effects in detail. Finally, the optimization problem of OSL task can be formulated as

$$\min_{\boldsymbol{D_V},\boldsymbol{D_A},\boldsymbol{D_H},\boldsymbol{C}} \varsigma_r + \lambda\varsigma_s + \delta\varsigma_\kappa. \quad (9)$$

Algorithm 1 Dictionary Learning

Require: Data sets $\boldsymbol{V}, \boldsymbol{A}$ and $\boldsymbol{H}$.

Ensure: Solutions $\boldsymbol{D}_V, \boldsymbol{D}_A$ and $\boldsymbol{D}_H$.

```
Initialize $\boldsymbol{D}_V, \boldsymbol{D}_A$ and $\boldsymbol{D}_H$.
While Not convergent do
    Fix $\boldsymbol{D}_V, \boldsymbol{D}_A$ and $\boldsymbol{D}_H$, update $\boldsymbol{C}$ by Eq. (11).
    Fix $\boldsymbol{C}$, update $\boldsymbol{D}_V, \boldsymbol{D}_A$ and $\boldsymbol{D}_H$ by Eq. (10).
end while
```

D. OPTIMIZATION ALGORITHM

In this subsection, the optimization algorithm will be respectively introduced. Obviously, the optimization function in (9) is hard to direct solve due to it is non-convex. Through the traditional dictionary learning method, as shown in algorithm 1, the optimal value can be approximated by multiple iterations. In each iteration, the optimal solution is updated by fixing dictionary and sparse coding respectively.

1) DICTIONARY UPDATING

This step renews each dictionaries $\boldsymbol{D}_V$, $\boldsymbol{D}_A$ and $\boldsymbol{D}_H$. When given the fixed spared code $\boldsymbol{C}^{(t)}$, the dictionaries $\boldsymbol{D}_V, \boldsymbol{D}_A$ and $\boldsymbol{D}_H$ can be update by

$$\{\boldsymbol{D}_V^{(t)}, \boldsymbol{D}_A^{(t)}, \boldsymbol{D}_H^{(t)}\} = \underset{\boldsymbol{D}_V, \boldsymbol{D}_A, \boldsymbol{D}_H}{argmin} \left\| \boldsymbol{V} - \boldsymbol{D}_V \boldsymbol{C}^{(t)} \right\|_F^2 + \left\| \boldsymbol{A} - \boldsymbol{D}_A \boldsymbol{C}^{(t)} \right\|_F^2 + \left\| \boldsymbol{H} - \boldsymbol{D}_H \boldsymbol{C}^{(t)} \right\|_F^2 \tag{10}$$

where t denotes the number of iterations. Combined with the design application, (10) can be converted to a simpler form

$$\{\boldsymbol{D}_V^{(t)}, \boldsymbol{D}_A^{(t)}, \boldsymbol{D}_H^{(t)}\} = \underset{\boldsymbol{D}_V, \boldsymbol{D}_A, \boldsymbol{D}_H}{argmin} \left\| \begin{bmatrix} \boldsymbol{V} \\ \boldsymbol{A} \\ \boldsymbol{H} \end{bmatrix} - \begin{bmatrix} \boldsymbol{D}_V \\ \boldsymbol{D}_A \\ \boldsymbol{D}_H \end{bmatrix} \boldsymbol{C}^{(t)} \right\|_F^2 . \tag{11}$$

which is helpful for data analysis, processing and programming. In this step, the optimization problem (11) can be solved by conventional dictionary learning algorithm. It can be easy solved by SPAMS [33].

2) COMMON SPARSE CODING UPDATING

This step renews the common sparse coding $\boldsymbol{C}$. When given the fixed dictionaries $\boldsymbol{D}_V^{(t)}, \boldsymbol{D}_A^{(t)}, D_H^{(t)}$, the $\boldsymbol{C}$ can be update by

$$\{\boldsymbol{C}^{(t+1)}\} = \min_{\boldsymbol{C}} \varsigma_r(\boldsymbol{D}_V^{(t)}, \boldsymbol{D}_A^{(t)}, \boldsymbol{D}_H^{(t)}) + \lambda\varsigma_s + \delta\varsigma_k. \tag{12}$$

The optimization problems are convex and can be easily solved by using CVXPY, which is an off-the-shelf Python-embedded modeling language for convex optimization problems [26], [27]. After practical experiments, the convergence condition will be satisfied after no more than ten iterations.

E. ONE-SHOT CLASSIFIER DESIGN

In this subsection, $\boldsymbol{D}_V^*$, $\boldsymbol{D}_A^*$, $\boldsymbol{D}_H^*$ and $\boldsymbol{C}^*$ respectively be denoted as the optimal dictionary solution and common sparse coding solution after multiple iterations procedure. The classifier is designed for the testing data set. For a new input testing sample i, its optimal multi-modal common sparse coding vector $\boldsymbol{x}^*$ can be obtained by

$$\boldsymbol{x}^* = \min_{\boldsymbol{x}} \left\| \begin{bmatrix} v \\ \boldsymbol{a} \\ \boldsymbol{h} \end{bmatrix} - \begin{bmatrix} \boldsymbol{D}_V^* \\ \boldsymbol{D}_A^* \\ \boldsymbol{D}_H^* \end{bmatrix} \boldsymbol{x} \right\|_2^2 \tag{13}$$

or

$$\boldsymbol{x}^* = \min_{\boldsymbol{x}} \left\| \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{a} \\ \boldsymbol{h} \end{bmatrix} - \begin{bmatrix} \boldsymbol{D}_V^* \\ \boldsymbol{D}_A^* \\ \boldsymbol{D}_H^* \end{bmatrix} \boldsymbol{x} \right\|_2^2 + \lambda \left|\boldsymbol{x}\right|_1 . \tag{14}$$

No matter how the ς_κ choose ς_e or ς_r, the label of testing sample can be decided by

$$l^* = \underset{l \in \{0,1,2,\ldots,N\}}{argmin} \frac{\boldsymbol{c}_l^* \boldsymbol{x}^*}{\left\|\boldsymbol{c}_l^*\right\|_2 \left\|\boldsymbol{x}^*\right\|_2} \tag{15}$$

where $\boldsymbol{x}^*$ can be obtained using (14), $\boldsymbol{c}_l^*$ is a child vector of the corresponding tag in $\boldsymbol{C}^*$, l^* is the label of input test sample.

IV. FEATURE REPRESENTATION

A. DATA SET

To explain our proposed model, the LMT data set is selected as the source of our test data. The LMT data set is briefly described in this section for further describing the approach we have proposed. The LMT data set is original development in [25]. The initial data set only contains 69 objects, with improved testing tools and methods, as shown in Fig. 3, the new data set contains 108 individual textures [28], which can be divided into nine material categories: 1) *Meshes*; 2) *Stones*; 3) *Glossy*; 4) *Wood Types*; 5) *Rubbers*; 6)*Fibers*; 7) *Foams*; 8)*Foils* and *Papers*; 9) *Textiles* and *Fabrics*.LMT-108 data set provides multi modalities data include acceleration, friction, image, metal detection, IR reflection and sound. In the process of collecting experimental data, one controls the handheld device, which is shown in Fig. 4, to slide freely over the material surface as he wishes, and ten sets of data were recorded at a time. In this paper, acceleration, image and sound is adopt to solve the OSL task. In LMT-108 data set, the acceleration is collected by a three-axes ADXL335 accelerometer (Analog Devices, Inc.) with a range of ±3g, sound is collected by CMP-MIC8, among all measurements, the sampling frequency of accelerometer is 10 KHz and that of sound is 44100 Hz, and size of image is 320×480.

B. DIMENSIONAL REDUCTION OF ACCELERATION

As a preliminary step of feature evaluation, it is necessary to reduce the dimension of acceleration data. By comparing different dimensionality reduction methods, SA321-z is selected as our dimensionality reduction method for acceleration data. The evaluated algorithms include:

1) *SoC321:* SoC321 is a computationally simple dimension reduction algorithm, which is to merely add up the measurements in different directions.

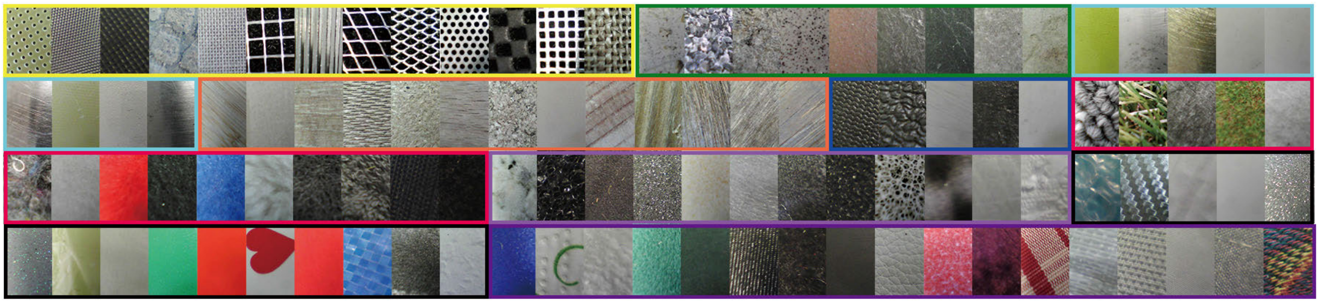


FIGURE 3. All textures. The samples are adopted from the LMT data set [28].

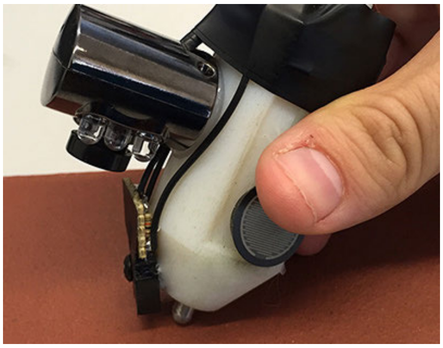


FIGURE 4. Texplorer overview. Copyright (2017) IEEE. Reprinted, with permission, from [28].

2) *Mag321:* Mag321 is one of the most common methods, which take the square root of the sum of the squares of the measurements in different directions. But obviously, because there is no negative value, some very valid information is missed.
3) *PCA321:* PCA is a widely used dimensionality reduction method, which can project 3d acceleration data onto one dimension.
4) *SA321-x/ SA321-y/ SA321-z:* SA321 is the simplest solution which uses a fixed axis. SA321-x, SA321-y and SA321-z represents the measurements on the x, y and z axes respectively.

Fig. 5 is the data of *G1EpoxyRasterPlate* after dimensional reduction of acceleration. Combined with our research content, in order to select the most suitable method, a set of evaluation methods are specially designed. The assessment process has the following three requirements:

1) KNN is employed as the classifier, the accuracy of KNN classification is the result of evaluation.
2) Only one sample of every class is selected as the training set, and the remaining nine instances are the testing set.
3) MFCC is adapted as the input features for the classifier.

The evaluation results are shown in table 1, Mag321, SA321-x and SA321-y are the worst performers, while PCA321 and SA321-z perform better. Table 1 shows that SA321-z have a distinct advantage than SA321-x, SA321-y. Maybe the reason is that, in actual measurements, the data in the z-axis direction is more sensitive to texture than the other two axes, retaining more reliable information, and the acceleration in the other two directions may more closely related to the force applied by the operator. To sum up, SA321-z is the best choice of all the methods discussed.

TABLE 1. Assessment of various dimension reduction method.

METHOD	Accuracy
SoC321	0.705
Mag321	0.684
PCA321	0.761
SA321-x	0.641
SA321-y	0.695
SA321-z	0.797

C. EVALUATION OF FEATURES

It is particularly important to select the appropriate features for each of its modal data, so some frequently-used features are calculated for the acceleration, image and sound data separately, and KNN classifier is used to simply evaluate in the case of OSL. The results are shown in table 2. Among all the features listed in table 2, PCA is adopted to reduce the dimension of those with a large number of features in order to facilitate operation. In our assessment, the features whose dimensionality is reduced included HOG, PSD and MLS. LBP has a best performance than others. For sound and acceleration, The MFCC stands out among many features, PNCC and PLP have worst performance.

V. EXPERIMENTAL RESULTS

In subsequent experimental description, all experimental results are based on LMT haptic texture database. The latest LMT data set provides 108 textures data which is introduced in the previous section in detail. In LMT-108 data set, ten samples are selected for each texture. It should be notified that only one sample of every textures is used as training set, and the rest is used as testing set, so there are 1 × 108 = 108 samples for training and 9 × 108 = 9072 samples for testing.

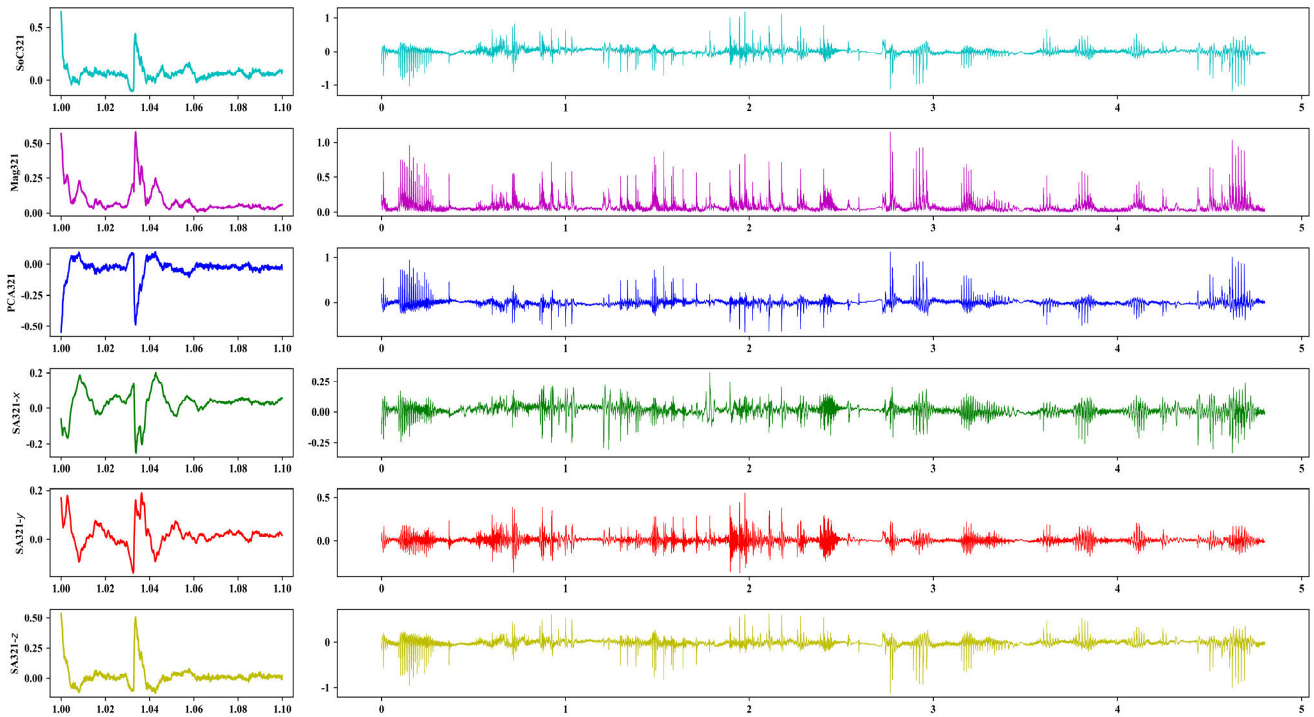


FIGURE 5. Dimensional reduction of acceleration.

TABLE 2. Assessment of vision, audio and acceleration features.

Modality	Feature	#	Accuracy
Image	*HOG*	40	0.30
	LBP	26	0.59
	GLCM	72	0.54
Sound	*MFCC*	30	0.76
	PSD	20	0.382
	PNCC	15	0.06
	LPC	13	0.241
	PLP	13	0.010
	MLS	20	0.43
Acc	*MFCC*	30	0.797
	PSD	20	0.209
	PNCC	15	0.041
	LPC	13	0.438
	PLP	13	0.019
	MLS	20	0.265

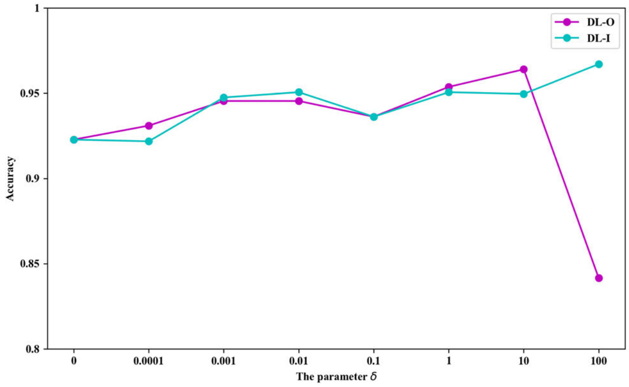


FIGURE 6. Recognition accuracy versus the regularization parameters δ. All experimental results are based on the same training set and test set.

After all parameters are set, the model of (1) is training for the texture recognition.

A. COMPARED METHOD

As we all know, there is no multi-modal fusion framework for image-acc-sound about OSL task. However, some common methods can be applied for this problem. Through experiments, several common methods are compared with our method. The methods used in our experiment include:

1) *Support vector machine (SVM)*: SVM is another frequently-used machine learning algorithm. It is a generalized linear classifier, based on supervised learning, that performs binary classification of data, and its decision boundary is a maximum-margin hyperplane that is used to solve learning samples. Multi-modal data can be simply fused to improve the recognition accuracy.
2) *Multi-layer perceptron (MLP)*: MLP is a common artificial neural network with forward structure, maps a set of input vectors to a set of output vectors. MLP is a generalization of perceptron, which overcomes the disadvantage that perceptron cannot recognize linear inseparable data.
3) *EasyMKL [34]*: EasyMKL algorithm can easily deal with hundreds of thousands of kernels and even more. EasyMKL uses only a limit amount of memory and it has only a linear time complexity.

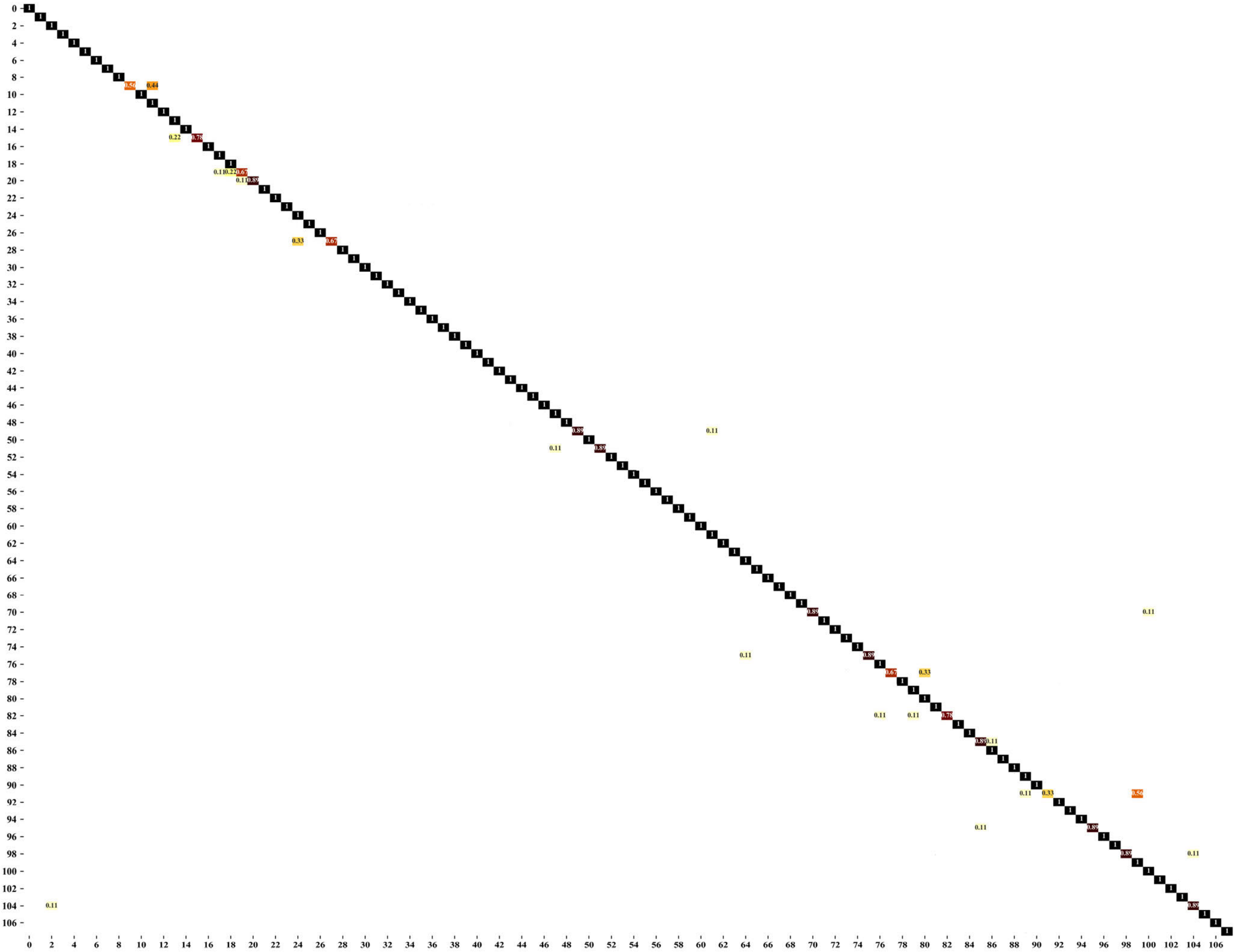


FIGURE 7. Confusion matrix of DL-I. Obviously, there were more misjudgments from *stones*, *Foils and Papers*, but *Wood* and *Fibers* have a high right rate of identification.

4) *DL-O:* It solved the proposed optimization problem (9), where ς_κ is set to ς_o, the (14) is chosen in this experiment.
5) *DL-I*: It solved the proposed optimization problem (9), where ς_κ is set to ς_i, the (14) is chosen in this experiment.

In order to evaluate the advantages and disadvantages of different methods, accuracy was chosen as the evaluation standard. By verifying the above methods in public available data set, the results are shown in table 3. Obviously, the accuracy of multi-modal fusion is far better than independent use, and both DL-O and DL-I are better than other methods, among which DL-I has the best performance due to its accuracy reaching 0.97, and it is possible that DL-O places too much emphasis on the sparsity of each row of the encoding matrix, resulting in the part of a matrix that has a larger value being focused on the encoding of a same sample data. The confusion matrix of DL-I with optimal parameters is shown in Fig. 7.

TABLE 3. Accuracy of methods.

METHOD	Accuracy
SVM	0.94
MLP	0.90
EasyKML	0.94
DL-O	0.96
DL-I	0.97

B. PERFORMANCE COMPARISON

To analysis the role of ς_κ and find out the latent optimal parameter, the value of parameter δ is gradually increased and then recording the result. As shown in Fig. 6, The following observations is obtained.

1) When δ is set to zero, the accuracy rate is about 0.89. When δ is not set as zero, the accuracy rate of both DL-O and DL-I increases, but excessive increase of the value of δ will have adverse effects.

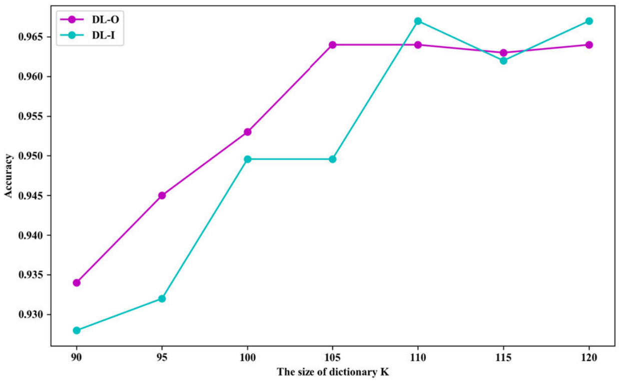


FIGURE 8. Recognition accuracy versus the size of dictionary K.

2) Obviously, DL-I has a better performance than DL-O, the accuracy rate and stability of DL-I are better than DL-O. The confusion matrix of DL-I is shown in Fig. 7 when δ is set to 1.5.
3) When δ is in the range of 0.06-0.09, the classification effect of DL-O is optimal. When δ is the range of 0.5-1.6, the classification effect of DL-I is optimal.
4) When the value of δ exceeds a certain range, the classification effect of DL-O and DL-I will decline to different degrees.

C. INFLUENCE OF DICTIONARY SIZE

Another important parameter of sparse dictionary learning is the dictionary size K, whose value has a direct impact on the recognition right rate. As shown in Fig. 8, as the size of the dictionary increases, the accuracy also increases, but when it reaches a certain limit, the accuracy decreases. As shown in Fig. 8, the accuracy of DL-I is better than that of DL-O.

D. MATERIALS COMPARISON

The accuracy of surface texture classification varies among different materials. LMT-108 data set comprises nine materials and different materials provide different numbers of textures. To explore the effect of our method on texture recognition in different materials, the effects of texture recognition is classified in different materials. First, DL-I was used to classify all the textures. Then, we counted the classification accuracy of textures based on different materials. The final results are shown in table 4, and it can be directly observed that when the material is wood types, our method achieves excellent results. For stones, the recognition rate of surface texture is relatively low.

E. COMPUTATIONAL EFFICIENCY ANALYSIS

For DL-O and DL-I, the time it takes to classify a testing set which contains 9072 samples is about 1.58 s. Like other dictionary learning methods, the cost of training time is long, and the method proposed by [18] can significantly improve the learning speed of the dictionary update stage, but for CVXPY, it still takes a long time to solve the convex optimization problem. During the process of dictionary learning, it takes about 8 s to iterate once. In our experiments, all method was implemented in Python 3.7 on a computer platform (3.2-GHz CPU and 8-G RAM).

TABLE 4. Accuracy of different materials.

MATERIAL	#	Accuracy
Meshes	13	0.966
Stones	9	0.926
Glossy	9	0.963
Wood	13	1.000
Rubbers	5	0.977
Fibers	15	0.993
Foams	12	0.981
Foils and Papers	15	0.956
Textiles and Fabrics	17	0.961

VI. CONCLUSION

In this paper, the OSL problem for texture recognition is faultlessly addressed by fusion various modalities. The proposed novel dictionary learning model not only perfectly fuses the various modalities measurements, but also simultaneously learns the latent common sparse code for the different modalities as well. Besides, in order to improve the recognition rate, a new regular term is proposed. Moreover, the iterative process of sparse dictionary learning is elaborated, and the dimensionality reduction method of 3d acceleration data proposed at present is elaborated based on practical application. In addition, the characteristics of each mode are evaluated to select the optimal term. In the experiments based on the public available data set, our method has showed a remarkable result with the best accuracy rate of 0.97.

ACKNOWLEDGMENT

The authors would like to thank Dr. M. Strese for the share of LMT haptic texture database.

REFERENCES

[1] G. V. L. de Lima, P. T. M. Saito, F. M. Lopes, and P. H. Bugatti, "Classification of texture based on bag-of-visual-words through complex networks," *Expert Syst. Appl.*, vol. 133, pp. 215–224, Nov. 2019, doi: 10.1016/j.eswa.2019.05.021.

[2] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi, "Deep filter banks for texture recognition, description, and segmentation," *Int. J. Comput. Vis.*, vol. 118, no. 1, pp. 65–94, May 2016.

[3] A. Song, Y. Han, H. Hu, and J. Li, "A novel texture sensor for fabric texture measurement and classification," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 7, pp. 1739–1747, Jul. 2014, doi: 10.1109/TIM.2013.2293812.

[4] M. M. Iskarous, H. H. Nguyen, L. E. Osborn, J. L. Betthauser, and N. V. Thakor, "Unsupervised learning and adaptive classification of neuromorphic tactile encoding of textures," in *Proc. BioCAS*, Cleveland, OH, USA, 2018, pp. 1–4.

[5] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[6] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. CVPR*, San Diego, CA, USA, Jun. 2005, pp. 539–546.

[7] B. P. Yuhas, M. H. Goldstein, Jr., and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Commun. Mag.*, vol. 27, no. 11, pp. 65–71, Nov. 1989.

[8] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Proc. ICME*, Singapore, 2010, pp. 1079–1084.

[9] M. Gurban, J. P. Thiran, T. Drugman, and T. Dutoit, "Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition," in *Proc. ICMI*, New York, NY, USA, 2008, pp. 237–240.

[10] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010, doi: 10.1007/s00530-010-0182-0.

[11] 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.

[12] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *Proc. ICME*, New York, NY, USA, 2000, pp. 1589–1592.

[13] W. H. Adams, G. Iyengar, C.-Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith, "Semantic indexing of multimedia content using visual, audio, and text cues," *EURASIP J. Adv. Signal Process.*, vol. 2003, Dec. 2003, Art. no. 987184, doi: 10.1155/S1110865703211173.

[14] J. Magalhães and S. Rüger, "Information-theoretic semantic multimedia indexing," in *Proc. CIVR*, Amsterdam, The Netherlands, 2007, pp. 619–626.

[15] S. B. Chaabane, M. Sayadi, F. Fnaiech, and E. Brassart, "Color image segmentation based on Dempster–Shafer evidence theory," in *Proc. IEEE MELECON*, Ajaccio, France, 2008, pp. 862–866.

[16] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 3, pp. 423–435, Mar. 2009.

[17] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, no. 2, pp. 349–396, Feb. 2003.

[18] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. ICML*, Montreal, QC, Canada, Jun. 2009, pp. 689–696.

[19] 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.

[20] H. Liu, F. Sun, B. Fang, and D. Guo, "Cross-modal zero-shot-learning for tactile object recognition," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published, doi: 10.1109/TSMC.2018.2818184.

[21] H. Liu, F. Sun, B. Fang, and S. Lu, "Multimodal measurements fusion for surface material categorization," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 2, pp. 246–256, Feb. 2018.

[22] F. Jeremy and L. Gerald, "Bayesian exploration for intelligent identification of textures," *Frontiers Neurorobotics*, vol. 6, p. 4, Jun. 2012, doi: 10.3389/fnbot.2012.00004.

[23] M. Strese, J. Lee, C. Schuwerk, Q. Han, H. Kim, and E. Steinbach, "A haptic texture database for tool-mediated texture recognition and classification," in *Proc. HAVE*, Richardson, TX, USA, 2014, pp. 118–123.

[24] Z. Kappassov, "Tactile sensing in dexterous robot hands—Review," *Robot. Auton. Syst.*, vol. 74, pp. 195–220, Dec. 2015, doi: 10.1016/j.robot.2015.07.015.

[25] M. Strese, C. Schuwerk, A. Iepure, and E. Steinbach, "Multimodal feature-based surface material classification," *IEEE Trans. Haptics*, vol. 10, no. 2, pp. 226–239, Apr./Jun. 2017.

[26] S. Diamond and S. Boyd, "A Python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 83, pp. 2909–2913, Apr. 2016.

[27] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd, "A rewriting system for convex optimization problems," *J. Control Decis.*, vol. 5, no. 1, pp. 42–60, 2018, doi: 10.1080/23307706.2017.1397554.

[28] C. M. Strese, Y. Boeck, and E. Steinbach, "Content-based surface material retrieval," in *Proc. WHC*, Munich, Germany, 2017, pp. 352–357.

[29] B. M. Lake, R. R. Salakhutdinov, and J. Tenenbaum, "One-shot learning by inverting a compositional causal process," in *Proc. NIPS*, Lake Tahoe, NV, USA, 2013, pp. 2526–2534.

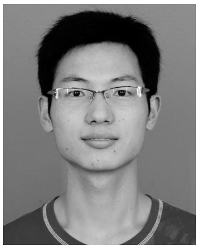
[30] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 1199–1208.

[31] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 3637–3645.

[32] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. ICLR*, Toulon, France, 2017, pp. 1–9.

[33] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 19–60, Jan. 2010.

[34] F. Aiolli and M. Donini, "EasyMKL: A scalable multiple kernel learning algorithm," *Neurocomputing*, vol. 169, pp. 215–224, Dec. 2015, doi: 10.1016/j.neucom.2014.11.078.



PENGWEN XIONG received the B.S. degree from the North University of China, in 2009, and the Ph.D. degree in instrument science and technology from Southeast University, China, in 2015. He visited the Laboratory for Computational Sensing and Robotics, Johns Hopkins University, from 2013 to 2014. He is currently an Associate Professor with the School of Information Engineering, Nanchang University, and holds a postdoctoral position at the School of Instrument Science and Engineering, Southeast University. His research interests include human–robot interaction, as well as robotic sensing and controlling.

KONGFEI HE received the B.E. degree in electrical engineering and automation from Chaohu University, China, in 2017. He is currently pursuing the master's degree in control science and engineering from Nanchang University, China.

AIGUO SONG received the B.S. degree in automatic control, in 1990, the M.S. degree in measurement and control from the Nanjing Aeronautics and Astronautics University, Nanjing, China, in 1993, and the Ph.D. degree in measurement and control from Southeast University, Nanjing, in 1996. From April 2003 to April 2004, he was a Visiting Scientist with the Lab for Intelligent Mechanical Systems, Northwestern University, Evanston, IL, USA. He is currently a Professor with the Department of Instrument Science and Engineering, Southeast University. His current research interests include haptic display and teleoperation robot.

PETER X. LIU (F'19) received the B.Sc. and M.Sc. degrees from Northern Jiaotong University, Beijing, China, in 1992 and 1995, respectively, and the Ph.D. degree from the University of Alberta, Edmonton, AB, Canada, in 2002. He has been with the Department of Systems and Computer Engineering, Carleton University, Canada, since July 2002, and is currently a Canada Research Professor. His research interests include interactive networked systems and teleoperation, robotics, intelligent systems, haptics, micromanipulation, context-aware intelligent networks, and their applications to biomedical engineering.