Multidim Syst Sign Process
DOI 10.1007/s11045-016-0389-0

Robotic grasping recognition using multi-modal deep extreme learning machine

Jie Wei[1,2] · **Huaping Liu**[1] · **Gaowei Yan**[2] · **Fuchun Sun**[1]

Received: 27 September 2015 / Revised: 22 January 2016 / Accepted: 19 February 2016

Abstract Recognizing which part of an object is graspable or not is important for intelligent robot to perform some complicated tasks. In order to obtain good grasping performance, learning rich representations efficiently from multi-modal RGB-D images is crucial. To address this problem, in this paper, we propose an effective multi-modal deep extreme learning machine structure. In this structure, unsupervised hierarchical extreme learning machine (ELM) is conducted for feature extraction for RGB and depth modalities separately. Then, the shared layer is developed by combining both RGB and depth features. Finally, the ELM is used as supervised feature classifier for final decision. Experimental validation on Cornell grasping dataset illustrates that the proposed multiple modality fusion method achieves better grasp recognition performance.

Keywords Robotic grasping · RGB-D multi-modal data · Deep ELM · Feature representations

1 Introduction

Robotic grasping is to grasp objects and possibly manipulate them by means of the gripper's fingers. Over the past decades, more and more researches (Bicchi and Kumar 2000; Hu et al. 2016; Saxena et al. 2008) are performed on the robotic grasping. In order to achieve grasping stability, task compatibility and operation effectiveness, several algorithms (Sahbani et al.

✉ Huaping Liu
hpliu@tsinghua.edu.cn

✉ Gaowei Yan
yangaowei@tyut.edu.cn

1 Department of Computer Science and Technology, Tsinghua University, State Key Laboratory of Intelligent Technology and Systems, TNLIST, Beijing, People's Republic of China

2 College of Information Engineering, Taiyuan University of Technology, Taiyuan, Shanxi, People's Republic of China

Published online: 03 March 2016

2012; Bohg et al. 2014) have been developed. Different approaches have been proposed to meet these goals, and substantial improvements have been claimed.

However, robotic grasping recognition is a challenging problem (Sahbani et al. 2012; Bohg et al. 2014). The goal of robotic grasping recognition lies in two aspects: (1) Infer the top locations where a robotic gripper could be placed. (2) Infer the optimal grasp for a given object that maximizes the chance of successfully grasping it. The two tasks are highly-coupled and therefore make the grasping recognition problem differ from the general object recognition problem (Lenz et al. 2015). In addition, this problem becomes more difficult when we are given the noisy image or partial view of the object from a camera.

Another significant challenge lies in the robotic grasping recognition is that the property of object cannot be well captured. In practice, the grasping performance not only depends on the pose and configuration of the robotic gripper, but also the shape and physical properties of the object to be grasped. The conventional camera usually provides limited RGB information about an object (Wang et al. 2013). This is insufficient when the object is irregular.

With the popularity of RGB-D Kinect cameras, the acquisition of depth information has made it easier to infer the optimal grasp for a given object beyond traditional RGB information. Therefore, a method to efficiently combine RGB information with depth information for robust image representation has become a core issue in RGB-D based Robotic Grasping Recognition.

In recent years, deep Learning has been a new area of machine learning research, which is about learning multiple levels of representation and abstraction that help to make sense of data such as images, sound, and text. In order to learn rich representations efficiently, deep learning has been applied to a lot of practical problems. For example, Convolutional Neural Network (CNN) is one kind of deep learning method, which can be applied to recognition of the speakers templates which are introduced to the system via a voice master card (Zaki et al. 1996).

Recently, more and more multi-modal deep learning methods take full advantages of the capacity of learning useful features. Wang et al. (2014) proposes an effective mapping mechanism based on stacked auto-encoders that enables seamless information retrieval from various types of media. Srivastava and Salakhutdinov (2012) proposes a multi-modal DBN architecture for learning a joint representation of image and text inputs. Ouyang et al. (2014) proposes to build a multi-source deep model in order to extract non-linear representation from three important information sources for human pose estimation. Wang et al. (2014) adapts the unsupervised feature learning technique as a multi-modality learning problem for RGB-D indoor scene labeling, in which higher-level features are derived and combined with lower-level features by stacking basic learning structure. Traditionally, the popular Gradient-Based learning algorithm used in deep learning methods is the back-propagation learning algorithm, in which all the hidden parameters in deep learning framework need to be fine-tuned multiple times. Thus, neural network cannot achieve good generalization performance at extremely fast learning speed (Huang et al. 2004).

Seen from the above analysis, the existing multi-modal deep learning methods cannot achieve excellent generalization performance with fast learning speed. Based on the fast learning speed and computational efficiency, ELM is more flexible and computationally attractive than traditional deep learning methods (Huang 2014). However, its shallow architecture makes it difficult to capture relevant higher-level abstractions (Cao et al. 2012; Cao and Lin 2015). In order to break this limitation, more and more deep ELM learning algorithm has been proposed. Cambria and Huang (2013) firstly introduces the ELM auto-encoder (ELM-AE), which represents features based on singular values. Yu et al. (2015) utilizes ELM as a base building block and incorporates random shift and kernelization as stacking elements.

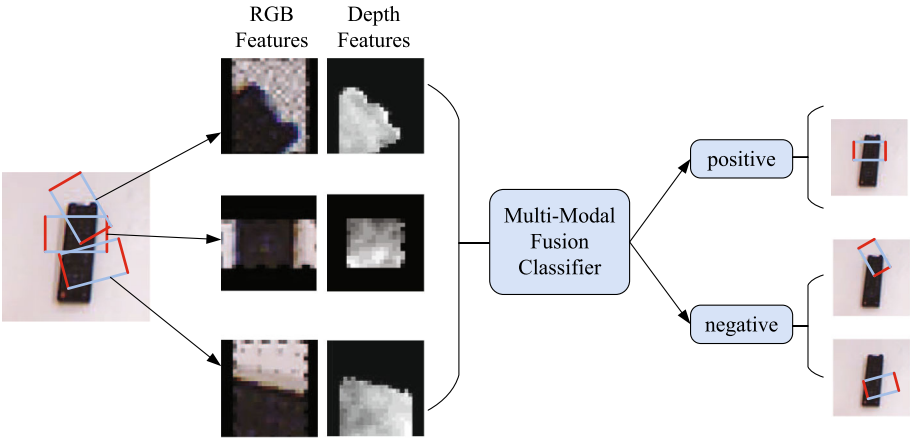

Fig. 1 Recognizing and evaluating each candidate rectangle

Zhu et al. (in press) employs the ELM-AE as the learning unit to learn local receptive fields at each layer. Uzair et al. (2015) proposes an efficient image set representation and the reconstruction error plays a role as a standard of classification. Tang et al. (2015) presents a new ELM sparse auto-encoder, which is utilized as the basic elements of H-ELM.

However, the aforementioned works do not refer to multi-modal problem (Yuan and Sun 2015). Thus, in this paper, we extend the deep ELM and propose a Multi-Modal Deep ELM-AE (MM-DELM) framework for robotic grasping. The proposed MM-DELM further improves the learning performance of the original ELM, while maintaining its advantages of training efficiency. The contributions of this work are summarized as follows:

1. We present a new multi-modal deep learning method under the framework of ELM. An important merit of such a method is that the training efficacy is highly improved.
2. The proposed multi-modal deep ELM method is used for deal with color and depth information which are captured for robotic grasping recognition.
3. We perform the experimental validation on the recently developed publicly available Cornell grasping dataset. The obtained results show that the proposed fusion method obtains rather promising results.

The remainder of this paper is organized as follows: Sect. 2 describes the problem of robotic grasping we would address; Sect. 3 introduces the related works, including the fundamental concepts and theories of ELM; Sect. 4 describes the proposed MM-DELM framework; Sect. 5 compares the performance of MM-DELM with single modality and one kind of shallow fusion framework on Cornell grasping dataset; while Sect. 6 concludes this paper.

2 Problem formulation

In the system for robotic grasping, as shown in Fig. 1, we first obtain an RGB-D image of the scene containing objects to be grasped from a Kinect mounted on the robot. Next, the corresponding raw features need to be extracted from color and depth images for each candidate grasp, which is searched over a large space of possible grasps. Figure 1 illustrates the RGB and depth features extracted from 3 potential grasps. Then these features are used as

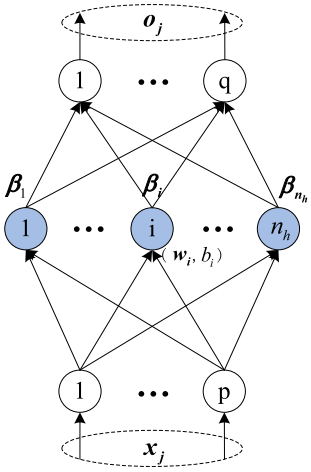

Fig. 2 The model of basic ELM

inputs to the Multi-Modal fusion classifier which scores each rectangle and evaluates whether robot could execute these grasps.

In this work, we will present an algorithm—the Multi-Modal Deep ELM-AE model as the multi-modal fusion classifier for robotic grasping recognition. In this part, multi-modal unsupervised hierarchical ELM is conducted for feature extraction. By doing so, it could achieve compact and meaningful feature representations; then the original ELM used as supervised feature classification is performed for final decision.

3 Extreme learning machines

The model of ELM proposed by Huang et al. (2004) is constituted by input layer, single-hidden layer, and output layer, as shown in Fig. 2. The aim of ELM is to solve above issues related with gradient-based algorithms. In ELM, input weights and single-hidden layer biases are arbitrarily chosen without iterative adjust, and the only parameters to be learned in training are the output weights which can be calculated by solving a single linear system (Huang et al. 2012). Therefore, ELM has been widely applied in regression and multi-class classification applications as an efficient learning algorithm (Cao et al. 2015; Akusok et al. 2015; Chen et al. 2015).

Different from traditional learning algorithms, given N training samples, $\{\mathbf{X}, \mathbf{T}\} = \{\mathbf{x}_j, \mathbf{t}_j\}_{j=1}^{N}$, where $\mathbf{x}_j \in \mathbf{R}^p$ and $\mathbf{t}_j \in \mathbf{R}^q$ are the jth input and target vectors respectively. The parameters p and q are the dimension of input and target vector respectively. To seek a regressor function from the input to the target, one popular form is the standard Single Hidden Layer Feed-forward network (SLFN) (Huang et al. 2006), where n_h single-hidden nodes fully connect the p input nodes to the q output nodes, which can be mathematically modeled as:

$$\mathbf{o}_j = \sum_{i=1}^{n_h} \boldsymbol{\beta}_i g(\mathbf{w}_i^T \mathbf{x}_j + b_i) = \mathbf{t}_j, \tag{1}$$

where $\mathbf{o}_j \in \mathbf{R}^q$ is the output vector of the jth training sample, $\mathbf{w}_i \in \mathbf{R}^p$ is the input weight vector connecting the input nodes to the ith hidden node, b_i is the bias of the ith hidden node, $g(\cdot)$ denotes hidden nodes nonlinear piecewise continuous activation functions.

The above N equations can be written compactly as:

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{T}, \tag{2}$$

where the matrix $\mathbf{T}$ is target matrix,

$$\mathbf{H} = \begin{bmatrix} g(\mathbf{w}_1^T \mathbf{x}_1 + b_1) & \cdots & g(\mathbf{w}_{n_h}^T \mathbf{x}_1 + b_{n_h}) \\ \vdots & \cdots & \vdots \\ g(\mathbf{w}_1^T \mathbf{x}_N + b_1) & \cdots & g(\mathbf{w}_{n_h}^T \mathbf{x}_N + b_{n_h}) \end{bmatrix}, \tag{3}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1^T \\ \vdots \\ \boldsymbol{\beta}_{n_h}^T \end{bmatrix}, \mathbf{T} = \begin{bmatrix} \mathbf{t}_1^T \\ \vdots \\ \mathbf{t}_N^T \end{bmatrix}. \tag{4}$$

The matrix $\mathbf{H}$ is the hidden layer output matrix, which can be randomly generated independent of the training data. $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_{n_h}]^T$ ($\boldsymbol{\beta}_i \in \mathbf{R}^q$) is the output weight matrix between the hidden nodes and the output nodes. Thus, training SLFNs simply amounts to getting the solution of a linear system (2) of output weights $\boldsymbol{\beta}$ (Feng et al. 2009).

According to the ELM Theory (see the related works referred in Feng et al. 2009), ELM aims to reach the smallest training error but also the smallest norm of weights:

$$\textit{Minimize} : \|\boldsymbol{\beta}\|_2^2 + \lambda \|\mathbf{H}\boldsymbol{\beta} - \mathbf{T}\|_2^2 . \tag{5}$$

A simple representation of the solution of the Eq. (5) is given explicitly by Huang et al. (2006) as

$$\widehat{\boldsymbol{\beta}} = \mathbf{H}^{\dagger}\mathbf{T}, \tag{6}$$

where $\mathbf{H}^{\dagger}$ is the Moore–Penrose generalized inverse of the hidden layer output matrix $\mathbf{H}$.

To improve generalization performance and make the solution more robust, we can add a regularization term (Ding et al. 2015). When the number of training samples is more than the number of hidden layer nodes, the output weight matrix $\boldsymbol{\beta}$ in regularized ELM can be expressed as:

$$\widehat{\boldsymbol{\beta}} = \left(\frac{\mathbf{I}}{\lambda} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{T}. \tag{7}$$

when the number of training samples is less than the number of hidden layer nodes, the output weight matrix $\boldsymbol{\beta}$ in regularized ELM can be expressed as:

$$\widehat{\boldsymbol{\beta}} = \mathbf{H}^T\left(\frac{\mathbf{I}}{\lambda} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{T}. \tag{8}$$

To summarize, ELM have two notably attractive features. Firstly, the learning speed of ELM is extremely fast. It can train SLFN much faster than classical learning algorithms. Secondly, all the parameters of the networks do not need to be tuned iteratively. Thus, the ELM tends to reach the solutions straightforward without the issue of over-fitting. These two features make ELM more flexible and attractive than traditional gradient-based algorithms.

However, It is crucial to learn rich representations efficiently, which could play an important role in achieving high generalization performance. Therefore, the heart of our approach is the process of the ELM-based unsupervised feature learning (Yu et al. 2015). Due to its shallow architecture, feature learning using ELM cannot capture relevant higher-level abstractions, even with a large number of hidden nodes. Although deep neural networks have been shown to yield good performance in various computer vision tasks, they are generally

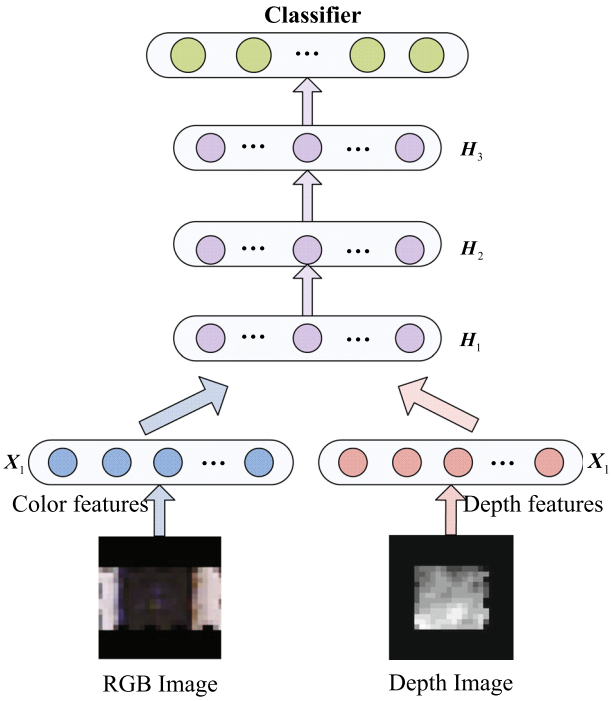

Fig. 3 The shallow fusion architecture

very slow in training. In order to make full use of advantages of these two methods, ELM-based Auto-Encoder (Deep ELM-AE) can be applied to extract the high level abstraction from the input data (Zhu et al. in press).

4 Multi-modal deep ELM-AE

As of today all kinds of multi-modal fusion method has been applied to various fields: in Jiang et al. (2012), MR image can be mapped onto the SPECT image in the corresponding area, in order to achieve early diagnosis of neurodegenerative diseases in the elderly. Porter and Liu (1994) explores a massively parallel neural array for multiple 2-D object recognition, in which a parallel modular form with each module being trained over a specific object class.

RGB-D image classification is also a multi-modality learning problem. Previous works (Lai et al. 2011; Bai and Wu 2014) have proved that combining RGB features and depth features together can dramatically increase the RGB-D based object recognition accuracy. Therefore, It is a core issue to find a method that effectively combines RGB information with depth information in RGB-D based image classification (Beksi and Papanikolopoulos 2015).

Jhuo et al. (2015) points out that there always exists a class of naive method which perform multi-modal fusion by combining RGB features and depth features as a concatenated vector that act as input of the framework, ignoring the particularity of information about specific modalities. In Fig. 3 we show such a architecture. From such a shallow fusion architecture we observe that the information sources with different statistical properties are mixed in the first hidden layer. Therefore, the performance cannot be expected to be satisfactory.

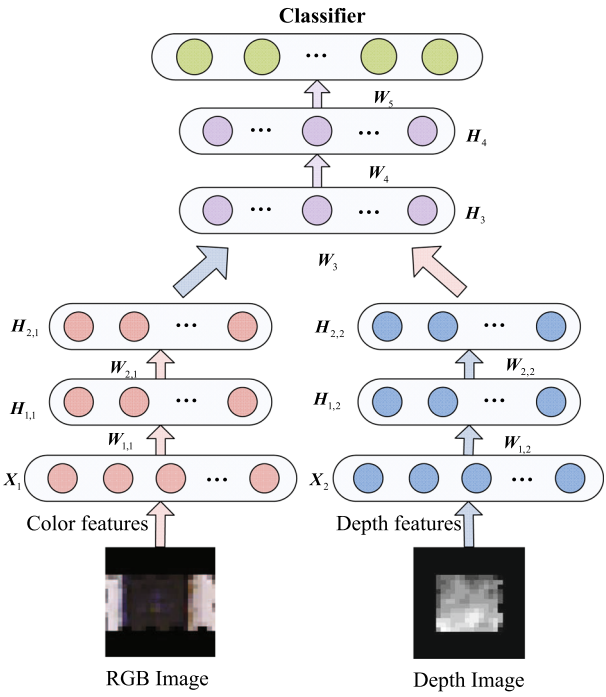

Fig. 4 The proposed multi-modal architecture

The approach proposed in this paper attempts to directly learn visual patterns from raw RGB-D data (Lai et al. 2011). To address this issue, a hierarchical learning framework, Multi-Modal Deep ELM-AE (MM-DELM), is proposed for robotic grasping recognition. The multi-modal training architecture is structurally divided into three separate phases: unsupervised feature representation for each modality separately, feature fusion representation and supervised feature classification, as shown in Fig. 4.

Formally, we consider a training set that contains m samples $(\mathbf{x}_1{}^{(i)}, \mathbf{x}_2{}^{(i)}, \mathbf{l}^{(i)})$, where $\mathbf{x}_1{}^{(i)} \in \mathbf{R}^{n_c}$ is the RGB feature vector, $\mathbf{x}_1{}^{(i)} \in \mathbf{R}^{n_d}$ is the depth feature vector, and $\mathbf{l}^{(i)} \in \mathbf{R}^2$ is label vector corresponding to the input $(\mathbf{x}_1{}^{(i)}, \mathbf{x}_2{}^{(i)})$. The parameters n_c and n_d are the dimension of RGB feature and depth feature. We define a new matrix $\mathbf{X}_1$ for the entire RGB feature set as the concatenation of the m training samples, $\mathbf{X}_2$ for the entire depth feature set, and $\mathbf{L}$ for the class label corresponding to the training samples. The MM-DELM classifier is learned through Algorithm 1.

Take Fig. 4, for instance. We perform feature learning to have high-level representations of each modality before they are mixed, in which RGB feature $\mathbf{X}_1$ and depth feature $\mathbf{X}_2$ are connected to two layers for constructing high level representation individually ($k_1 = 2$). Mathematically, starting from $\mathbf{H}_{0,j} = \mathbf{X}_j (j = 1, 2)$ for the first layer as our initial input, the output of the two hidden layer can be separately represented as:

$$\mathbf{H}_{i,j} = g\left(\mathbf{W}_{i,j}\mathbf{H}_{i-1,j} + \mathbf{B}_{i,j}\right), \textit{for } i, j = 1, 2, \tag{9}$$

where $g(\cdot)$ is activation function and we choose the sigmoid function. $\mathbf{H}_*$($*$ can represent that the hidden nodes belong to which layer and modality) is hidden layer matrix representing non-linear representations extracted from RGB features $\mathbf{X}_1$ or depth features $\mathbf{X}_2$. For example,

Algorithm 1 robotic grasping recognition through Multi-Modal Deep ELM-AE

Input: RGB feature matrix $\mathbf{X}_1 = [\mathbf{x}_1{}^{(1)}, \mathbf{x}_1{}^{(2)}, \ldots, \mathbf{x}_1{}^{(m)}]^T$ depth feature vector $\mathbf{X}_2 = [\mathbf{x}_2{}^{(1)}, \mathbf{x}_2{}^{(2)}, \ldots, \mathbf{x}_2{}^{(m)}]^T$ and label matrix $\mathbf{L} = [\mathbf{l}_1{}^{(1)}, \mathbf{l}_1{}^{(2)}, \ldots, \mathbf{l}_1{}^{(m)}]^T$ corresponding to each training instance, $\mathbf{x}_1^{(i)} \in \mathbf{R}^{n_c}$, $\mathbf{x}_2^{(i)} \in \mathbf{R}^{n_d}$, $\mathbf{l}^{(i)} \in \mathbf{R}^2$
Output: weight matrixes $\mathbf{W}_{i,j}$ for $i \in [1, k_1]$, $j \in [1, 2]$, and $\mathbf{W}_{s+1}$ for $s \in [k_1, k_1 + k_2 - 1]$
Initialization: Choose separate modality depth k_1, fusion learning model depth k_2, the node number n_h and the regularization term λ_h for each hidden layer.
for $j = 1$ to 2 do
  $\mathbf{H}_{0,j} = \mathbf{X}_j$
  for $i = 1$ to k_1 do
    Randomly generate hidden input weight matrix $\mathbf{w}_{i,j}$, bias matrix $\mathbf{B}_{i,j}$;
    Compute hidden layer output $\mathbf{H}_{i,j} = g\left(\mathbf{w}_{i,j}\mathbf{H}_{i-1,j} + \mathbf{B}_{i,j}\right)$;
    Calculate $\widehat{\boldsymbol{\beta}}_{i,j}$ using the equation (7), (8) or (13) with $\mathbf{H} = \mathbf{H}_{i,j}$, $\mathbf{T} = \mathbf{H}_{i-1,j}$;
    Compute $\mathbf{W}_{i,j}$ by equation (14): $\mathbf{W}_{i,j} = \widehat{\boldsymbol{\beta}}_{i,j}^T$;
    Update $\mathbf{H}_{i,j} = g\left(\mathbf{W}_{i,j}\mathbf{H}_{i-1,j} + \mathbf{B}_{i,j}\right)$;
  end for
end for
$\mathbf{H}_{k_1} = [\,\mathbf{H}_{k_1,1}^T\ \mathbf{H}_{k_1,2}^T\,]^T$
for $s = k_1$ to $k_1 + k_2 - 1$ do
  Randomly generate hidden input weight matrix $\mathbf{w}_{s+1}$, bias matrix $\mathbf{B}_{s+1}$;
  Compute hidden layer output $\mathbf{H}_{s+1} = g\left(\mathbf{w}_{s+1}\mathbf{H}_s + \mathbf{B}_{s+1}\right)$;
  Calculate $\widehat{\boldsymbol{\beta}}_{s+1}$ using the equation (7), (8) or (13) with $\mathbf{H} = \mathbf{H}_{s+1}$, $\mathbf{T} = \mathbf{H}_s$;
  Compute $\mathbf{W}_{s+1}$ by equation (14): $\mathbf{W}_{s+1} = \widehat{\boldsymbol{\beta}}_{s+1}^T$;
  Update $\mathbf{H}_{s+1} = g\left(\mathbf{W}_{s+1}\mathbf{H}_s + \mathbf{B}_{s+1}\right)$;
end for
Calculate $\widehat{\boldsymbol{\beta}}_{k_1+k_2+1}$ using the equation (7), (8) or (13) with $\mathbf{H} = \mathbf{H}_{k_1+k_2}$, $\mathbf{T} = \mathbf{L}$;
Compute $\mathbf{W}_{k_1+k_2+1}$ by equation (15): $\mathbf{W}_{k_1+k_2+1} = \widehat{\boldsymbol{\beta}}_{k_1+k_2+1}$;

$\mathbf{H}_{i,j}$ represents the ith layer feature representation of jth modality. In the same way, $\mathbf{B}_*$ is the bias matrix of the corresponding hidden layer and modality.

These high level representations of different information sources—$\mathbf{H}_{2,1}$, $\mathbf{H}_{2,2}$ are mixed in a two-layer stacked structure ($k_2 = 2$) to get well joint representation $\mathbf{H}_4$. At the inference stage, the combination process is as follows:

$$\mathbf{H}_2 = \left[\,\mathbf{H}_{2,1}^T\ \mathbf{H}_{2,2}^T\,\right]^T, \tag{10}$$

$$\mathbf{H}_s = g(\mathbf{W}_s\mathbf{H}_{s-1} + \mathbf{B}_s),\ for\ s = 3, 4. \tag{11}$$

Finally, the original ELM is performed to make final decision based on the joint representation:

$$\widehat{\mathbf{T}} = g(\mathbf{W}_5\mathbf{H}_4). \tag{12}$$

Through the proposed approach, multi-modal system can be developed as one whole system rather than being developed as separate expert systems for each modality.

Here, we consider a fully connected multi-modal multi-layer network with $k_1 + k_2 = 4$ hidden layers. Let $\mathbf{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_{k_1+k_2+1}\}$ ($\mathbf{W}_i = \{\mathbf{W}_{i,1}, \mathbf{W}_{i,2}\}$, $i = 1, \ldots, k_1$) denotes the parameters of the network that need to be learned. In our paper, the Deep ELM-AE is applied to learning the parameters $\mathbf{W}$.

The Deep ELM-AE is designed by using the encoded outputs to approximate the original inputs by minimizing the reconstruction errors (Cambria and Huang 2013; Tang et al. 2015). This architecture can find higher-level representations, thus can potentially capture relevant higher-level abstractions.

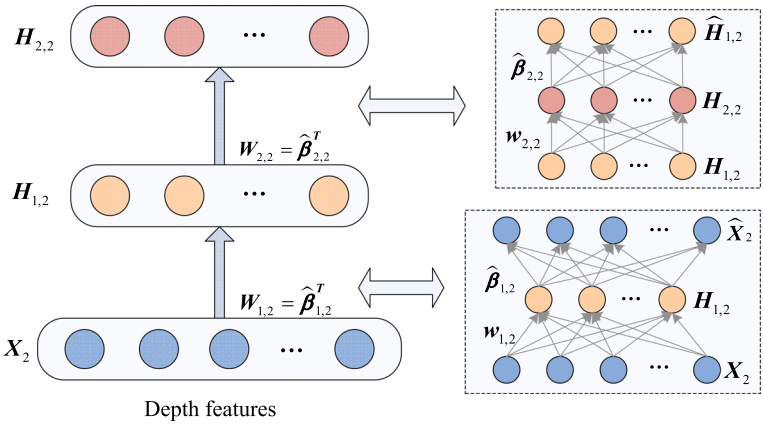

Fig. 5 Detailed illustration of the DELM representation learning

Figure 5 illustrates the process of learning representation from the depth features $\mathbf{X}_2$, which is similar to the process of color features learning and fusion learning. To simplify training, each hidden layer of Deep ELM-AE is an independent ELM, and functions as a separated feature extractor, whose target is same as its input. By minimizing the reconstruction error, we get the lower dimensional representation of the input data, which is regarded as input in the next layer. For instance, $\widehat{\boldsymbol{\beta}}$ is learned by considering a corresponding ELM with target matrix $\mathbf{T} = \mathbf{X}_2$, and the calculated $\mathbf{H}_{1,2}$ serves as the input of the second ELM.

Thus, the system becomes a linear model and the output weights $\boldsymbol{\beta}$ can be analytically determined by the Eq. (7) or (8) depending on the number of nodes in the hidden layer. Note that, when the number of nodes between two consecutive layers is equal, the output weights $\boldsymbol{\beta}$ are calculated as the solution to the orthogonal procrustes problem (Uzair et al. 2015). To find the orthogonal matrix $\boldsymbol{\beta}$, we use the singular value decomposition $\mathbf{M}$ to compute $\boldsymbol{\beta}$ as following equation:

$$\begin{aligned} \mathbf{M} &= \mathbf{H}^T\mathbf{T}, \\ \mathbf{M} &= \mathbf{U}\sum\mathbf{V}^T, \\ \widehat{\boldsymbol{\beta}} &= \mathbf{U}\mathbf{V}^T. \end{aligned} \tag{13}$$

It is important to note that the weight vectors connecting the input layer to each unit of the first hidden layer need to be orthonormal to each other, automatically learning the non-linear structure of data in a very efficient manner.

By this way, we can get the parameters $\widehat{\boldsymbol{\beta}}_k$ that re-projects the lower dimensional representation of the input data back to its original space while minimizing the reconstruction error. Therefore, this projection matrix is data-driven and hence used as the weights of the kth layer.

$$\mathbf{W}_k = \widehat{\boldsymbol{\beta}}_k^T \ (k = 1, \ldots, k_1 + k_2). \tag{14}$$

Finally, the learned features $\mathbf{H}_{k_1+k_2}$ are transferred to the original ELM to model the mapping between feature representation and the label. The $\mathbf{H}_{k_1+k_2}$ can be regarded as the hidden layer of the original ELM, in which $\boldsymbol{\beta}_{k_1+k_2+1}$ can be obtained easily. Thus, we can get the parameter $\mathbf{W}_{k_1+k_2+1}$ as following equation:

$$\mathbf{W}_{k_1+k_2+1} = \widehat{\boldsymbol{\beta}}_{k_1+k_2+1}. \tag{15}$$

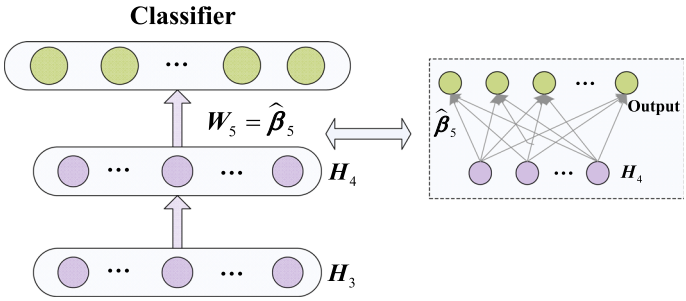

Fig. 6 The original ELM makes the final decision

In MM-DELM, RGB information and depth information are combined in an effective way, and particularity and information for specific modalities have been learned; By using the calculated $\boldsymbol{\beta}$ as the weights of the proposed auto-encoder, the inner product of the inputs and learned features would reflect the compact representations of the input in a very efficient manner; In contrast to deep networks, MM-DELM also does not require expensive iterative fine tuning of the weights (Fig. 6).

5 Experimental result

5.1 Dataset

We used the Cornell grasping dataset (Lenz et al. 2015) for our experiments, which is available at http://pr.cs.cornell.edu/deepgrasping. This dataset contains 1035 images of 280 graspable objects, several of which are shown in Fig. 7. Each image is labeled with roughly equal numbers of ground-truth positive and negative grasping rectangles, representing these rectangles are graspable or non-graspable. Some examples of these rectangles are shown in Fig. 1. By doing this, the deep ELM-AE can learn good representations for graspable rectangles even from non-graspable data.

Our algorithm uses only local information—specifically, we extract the RGB-D subimages contained within each rectangle, and use these to generate features for that rectangle. In this section, we will define the set of raw features which our algorithm will use, forming color features $\mathbf{X}_1$ and depth features $\mathbf{X}_2$ in the equations above.

The color features are extracted from RGB's three 24×24 pixel channels, giving $24 \times 24 \times 3 = 1728$ input features. The three channels are the image in color space, used because it represents image intensity and color separately. The depth features simply contain the depth channel of the image, giving $24 \times 24 = 576$ input features. They are computed after the image is aligned to the gripper so that they are always relative to the gripper plates.

5.2 Result

We compare our algorithm in the Cornell grasping dataset with other single-modality networks trained in a similar manner where two separate sets of first layer features are learned for the depth channel, the combination of the RGB channels. A kind of multi-modal learning method is also compared, in which color features and depth features are combined in a simple way, as shown in the Fig. 3. In order to apply Deep ELM-based Auto-Encoder to

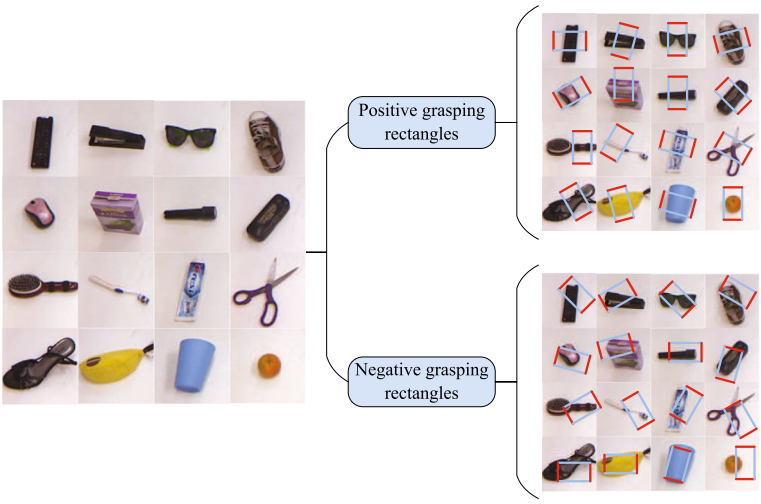

Fig. 7 Example objects from the Cornell grasping dataset

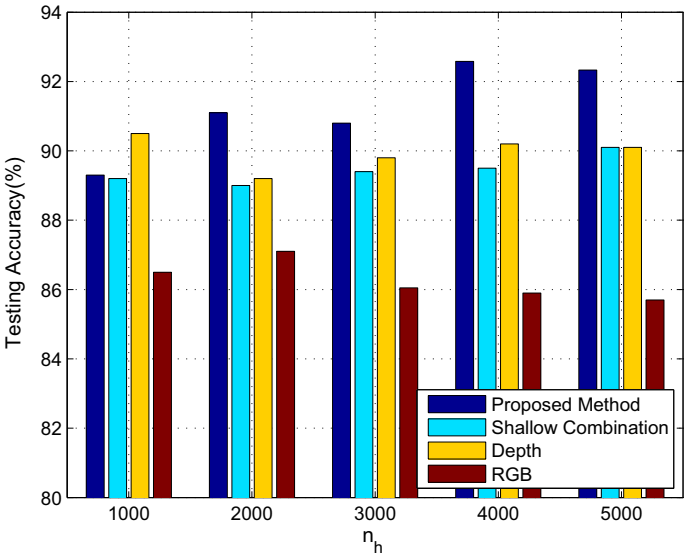

Fig. 8 The testing accuracy versus the number of hidden layer nodes

multi-modal data, the simple multi-modal method combines RGB features and depth features as a concatenated vector that act as input of the framework (Fig. 8).

Table 1 shows that the Multi-Modal Deep ELM-AE outperforms other methods for recognition, when the parameter $\lambda = 1$, $n_h = 4000$. Compared with single-modality networks, our Multi-Modal Deep ELM-AE is able to learn rich representations efficiently. And compared with RGB features, depth features can obtain the higher performance.

From Table 1 we observe that the multi-modal fusion method achieves the best scores. This is not surprising because combining both modality information usually leads to more robust results than either modality alone. In practice, depth image usually provides noisy

Table 1 Recognition results for different modalities

Modality	Accuracy (%)
RGB	87.47
Depth	90.52
Shallow fusion	90.39
Proposed method	92.58

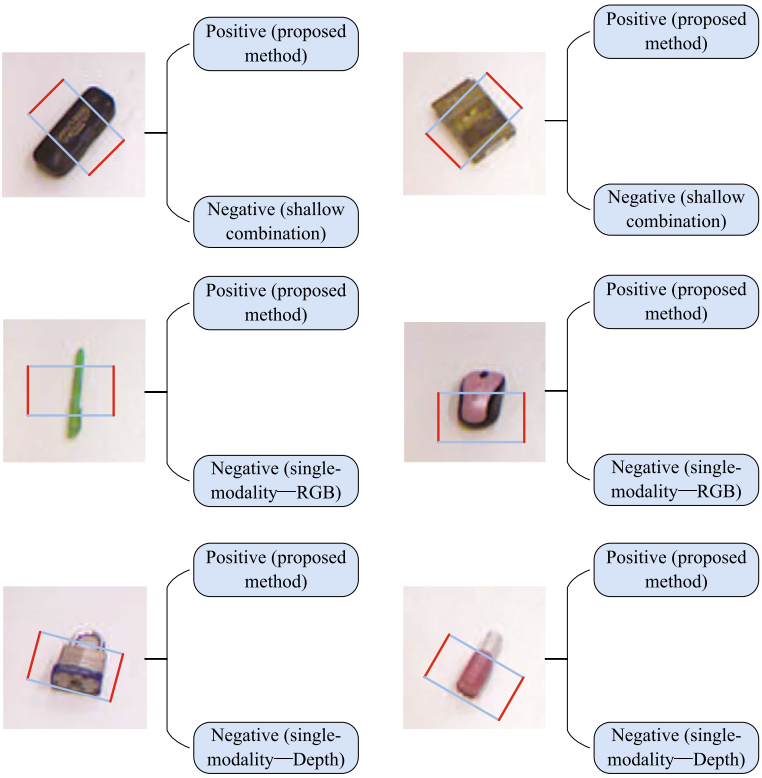

Fig. 9 Comparison for some potential grasping rectangles

information, but it can provides very useful information for some special cases, such as rims of monochromatic objects. As a result, integrating multi-modal information indeed plays important roles in recognizing good robotic grasps.

The proposed method and the modality of shallow fusion are both integrating multi-modal information, but the proposed method outperforms the latter, as shown in the Fig. 9, because different statistical properties are mixed in the the modality of shallow fusion ignoring the particularity of information about specific modalities.

Figure 9 illustrates some potential grasping rectangles, which need to be determined whether these grasp can be executed. As shown in Fig. 9, compared with shallow fusion and two single-modal methods, the proposed method always obtains a precise judgement simply.

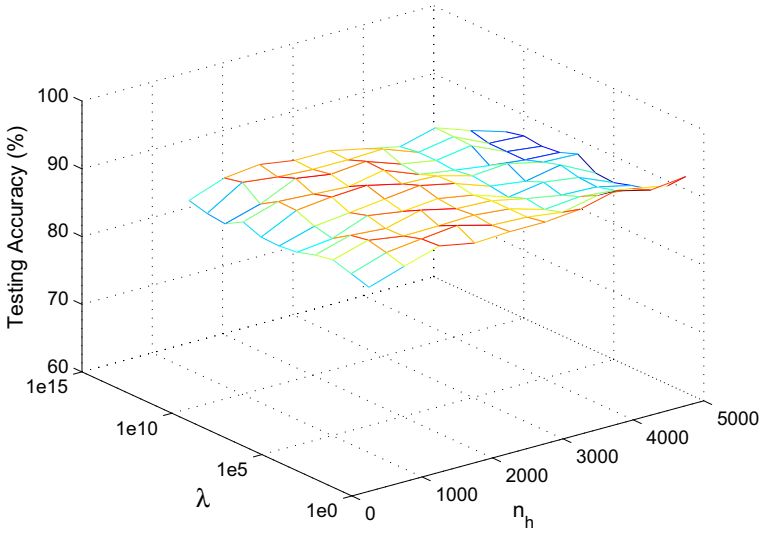

Fig. 10 Testing accuracy of MM-DELM in terms of n_h and λ

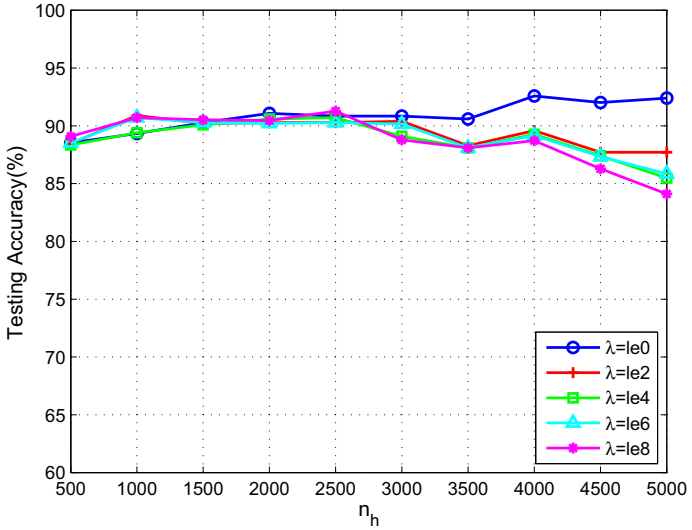

Fig. 11 Testing accuracy of MM-DELM in terms of n_h

5.3 Parameter sensitivity analysis

To analyse the roles of these parameters, we perform the sensitivity analysis. The most important two parameters in the proposed MM-DELM include the parameter λ for the regularized least mean square calculation, and the number of hidden nodes n_h. Therefore, as shown in Fig. 10, we vary the value of λ within the set $\{10^0, 10^1, \ldots, 10^9, 10^{10}\}$, and the value of n_h within the set $\{500, 1000, \ldots, 5000\}$ to analyze the performance variations.

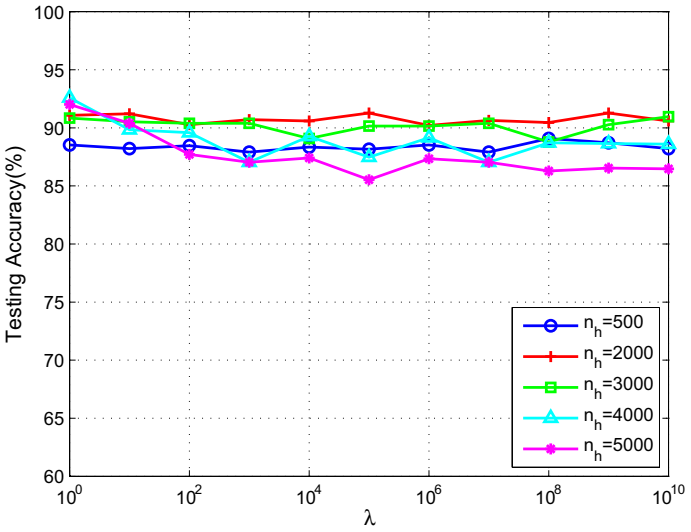

Fig. 12 Testing accuracy of MM-DELM in terms of λ

In order to analyse the sensitivity of recognition result in terms of the two parameters separately, we do the following two experiments independently. Figure 11 depicts the learning accuracies of MM-DELM in the n_h subspace, where the parameter λ is prefixed. Meanwhile, Fig. 12 prefixes the parameter n_h and illustrates the testing accuracy of MM-DELM in the λ subspace. It can be seen that MM-DELM follows a similar convergence property of ELM, and the performances tend to be quite stable in a wide range of n_h and λ. So we can choose the best parameters in the experiment according to the principle of having a high accuracy of test set like the overwhelming majority of the related work (Tang et al. 2015; Huang et al. 2012).

6 Conclusion

In this paper, we have proposed a novel multi-modal training scheme MM-DELM based on raw RGB-D images for robotic grasping recognition, in which particularity of information about RGB and depth modalities has been learned and combined in an effective way. In this structure, MM-DELM takes full advantage of the hierarchical ELM to learn the high level representation from RGB-D multi-modal data. As an artificial neural network, DELM takes advantages of both deep learning and extreme learning machine. DELM can approximate the complicated function without iterative fine-tuning. Thus, the proposed method could obtain more robust and better performance and it could be more flexible and computationally attractive than traditional deep learning methods. We also verified the generality and capability of MM-DELM on the Cornell grasping dataset. Compared with the single-modal and shallow fusion method, the training of MM-DELM is much faster and achieves higher learning accuracy.

Acknowledgments This work was supported in part by the National Key Project for Basic Research of China under Grant 2013CB329403; in part by National High-tech Research and Development Plan under

Grant 2015AA042306; in part by the National Natural Science Foundation of China under Grants 61210013 and 61450011; and in part by the Tsinghua University Initiative Scientific Research Program under Grant 20131089295.

References

Akusok, A., Miche, Y., Karhunen, J., Bjork, K. M., Nian, R., & Lendasse, A. (2015). Arbitrary category classification of websites based on image content. *IEEE Computational Intelligence Magazine*, *10*(2), 30–41.

Bai, J., & Wu, Y. (2014). SAE-RNN deep learning for RGB-D based object recognition. In *Intelligent computing theory*. Lecture notes in computer science, Vol. 8588, pp. 235–240.

Beksi, W. J., & Papanikolopoulos, N. (2015). Object classification using dictionary learning and RGB-D covariance descriptors. In *International conference on robotics and automation (ICRA)* (pp. 1–6).

Bicchi, A., & Kumar, V. (2000). Robotic grasping and contact: A review. In *International conference on robotics and automation (ICRA)* (pp. 348–353).

Bohg, J., Morales, A., Asfour, T., & Kragic, D. (2014). Data-driven grasp synthesis—A survey. *IEEE Transactions on Robotics*, *30*(2), 289–309.

Cambria, E., & Huang, G. (2013). Extreme learning machines-representational learning with ELMs for big data. *IEEE Intelligent Systems*, *28*(6), 30–59.

Cao, J. W., Chen, T., & Fan, J. Y. (2015). Landmark recognition with compact BoW histogram and ensemble ELM. *Multimedia Tools and Applications*. doi:10.1007/s11042-014-2424-1.

Cao, J., & Lin, Z. (2015). Extreme learning machine on high dimensional and large data applications: A survey. *Mathematical Problems in Engineering*. doi:10.1155/2015/103796.

Cao, J., Lin, Z., Huang, G.-B., & Liu, N. (2012). Voting based extreme learning machine. *Information Sciences*, *185*(1), 66–77.

Chen, Y., Yao, E., & Basu, A. (2015). A 128 channel extreme learning machine based neural decoder for brain machine interfaces. *IEEE Transactions on Biomedical Circuits and Systems* (in press).

Ding, S., Zhang, N., Xu, X., Guo, L., & Zhang, J. (2015). Deep extreme learning machine and its application in EEG classification. *Mathematical Problems in Engineering*. doi:10.1155/2015/129021.

Feng, G., Huang, G., Lin, Q., & Gay, R. (2009). Error minimized extreme learning machine with growth of hidden nodes and incremental learning. *IEEE Transactions on Neural Networks*, *20*(8), 1352–1357.

Huang, G., Zhu, Q., & Siew, C. (2004). Extreme learning machine: A new learning scheme of feedforward neural networks. In *Proceedings of international joint conference on neural network (IJCNN)* (Vol. 2, pp. 985–990).

Huang, G. B. (2014). An insight into extreme learning machines: Random neurons, random features and kernels. *Cognitive Computation*, *61*(1), 376–390.

Huang, G., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transaction on Systems, Man, and Cybernetics, Part B: Cybernetics*, *42*(2), 513–529.

Huang, G., Zhu, Q., & Siew, C. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, *70*, 489–501.

Hu, X., Zhang, X., Liu, M., Chen, Y., Li, P., Liu, J., et al. (2016). High precision intelligent flexible grasping front-end with CMOS interface for robots application. *Science China Information Sciences*, *59*, 032203(11).

Jhuo, I. H., Gao, S., Zhuang, L., & Lee, D. T. (2015). Unsupervised feature learning for RGB-D image classification. In *Asian conference on computer vision (ACCV)* (pp. 276–289).

Jiang, C. F., Chang, C. C., & Huang, S. H. (2012). Regions of interest extraction from SPECT images for neural degeneration assessment using multimodality image fusion. *Multidimensional Systems and Signal Processing*, *23*(4), 437–449.

Lai, K., Bo, L., Ren, X., & Fox, D. (2011). A large-scale hierarchical multi-view RGB-D object dataset. In *International conference on robotics and automation (ICRA)* (pp. 1817–1824).

Lenz, I., Lee, H., & Saxena, A. (2015). Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, *34*(4–5), 705–724.

Ouyang, W., Chu, X., & Wang, X. (2014). Multi-source deep learning for human pose estimation. In *Computer vision and pattern recognition (CVPR)* (pp. 2337–2344).

Porter, W. A., & Liu, W. (1994). Object recognition by a massively parallel 2-D neural architecture. *Multidimensional Systems and Signal Processing*, *5*(2), 179–201.

Sahbani, A., El-Khoury, S., & Bidaud, P. (2012). An overview of 3D object grasp synthesis algorithms. *Robotics and Autonomous Systems*, *60*, 326–336.

Saxena, A., Driemeyer, J., & Ng, A. Y. (2008). Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, *27*(2), 157–173.
Srivastava, N., & Salakhutdinov, R. (2012). Learning representations for multi-modal data with deep belief nets. In *International conference on machine learning workshop* (pp. 1–8).
Tang, J., Deng, C., & Huang, G. (2015). Extreme learning machine for multilayer perceptron. *IEEE Transactions on Neural Networks and Learning Systems*. doi:10.1109/TNNLS.2015.2424995.
Uzair, M., Shafait, F., Ghanem, B., & Mian, A. (2015). Representation learning with deep extreme learning machines for efficient image set classification. arXiv preprint arXiv:1503.02445, pp. 1–10.
Wang, A., Lu, J., Wang, G., Cai, J., & Cham, T. J. (2014). Multimodal unsupervised feature learning for RGB-D scene labeling. In *European conference on computer vision (ECCV)* (pp. 453–467).
Wang, W., Ooi, B. C., Yang, X., Zhang, D., & Zhuang, Y. (2014). Effective multi-modal retrieval based on stacked auto-encoders. *Proceedings of the VLDB Endowment*, *7*(8), 649–660.
Wang, J., Su, G., Xiong, Y., Chen, J., Shang, Y., Liu, J., et al. (2013). Sparse representation for face recognition based on constraint sampling and face alignment. *Tsinghua Science and Technology*, *1*, 62–67.
Yuan, Y., & Sun, F. (2015). Data fusion-based resilient control system under DoS attacks: A game theoretic approach. *International Journal of Control Automation and Systems*, *13*(3), 513–520.
Yu, W., Zhuang, F., He, Q., & Shi, Z. (2015). Learning deep representations via extreme learning machines. *Neurocomputing*, *149*, 308–315.
Zaki, M., Ghalwash, A., & Elkouny, A. A. (1996). CNN: A speaker recognition system using a cascaded neural network. *Multidimensional Systems and Signal Processing*, *7*(1), 87–99.
Zhu, W., Miao, J., Qing, L., & Huang, G. (in press). Hierarchical extreme learning machine for unsupervised representation learning. *Neurocomputing*.

Jie Wei received the B.Eng. degree from Taiyuan University of Technology, Taiyuan, China, in 2013. She is studying for the M.Eng. degree in Control Science and Engineering from the College of Information Engineering, Taiyuan University of Technology. Her research interests include deep learning and the technology of multi-modal fusion.

Huaping Liu is an associate professor in Department of Computer Science and Technology, Tsinghua University. He serves as Associate Editor of some journals including IEEE Robotics & Automation Letters, Neurocomputing, International Journal of Control, Automation and Systems, and some conferences including ICRA and IROS. His research interests include robot perception and learning.

Gaowei Yan received his B.Eng., M.Eng., and Dr.Eng. degrees all from Taiyuan University of Technology, Taiyuan, China, in 1992, 2003, 2007, respectively. In 1992, he joined the Department of Automation in Taiyuan University of Technology, where he is currently a professor of College of Information Engineering in Taiyuan University of Technology. His major research interests include soft sensor, machine learning, deep learning and multi-sensor fusion.

Fuchun Sun is a full professor in Department of Computer Science and Technology, Tsinghua University. He is the recipient of National Science Fund for Distinguished Young Scholars. He serves as Associate Editor of a series of international journals including IEEE TRANSACTIONS ON FUZZY SYSTEMS, Mechatronics, Robotics and Autonomous Systems. His research interests include intelligent control and robotics.