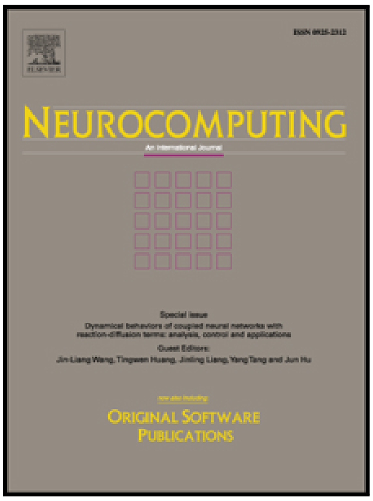

ISSN 0925-2312
NEUROCOMPUTING
An International Journal
Special issue
Dynamical behaviors of coupled neural networks with reaction-diffusion terms: analysis, control and applications
Guest Editors:
Jin-Liang Wang, Tingwen Huang, Jinling Liang, Yang Tang and Jun Hu
now also including
ORIGINAL SOFTWARE PUBLICATIONS

Multi-Modal Local Receptive Field Extreme Learning Machine for Object Recognition

Huaping Liu, Fengxue Li, Xinying Xu, Fuchun Sun

PII: S0925-2312(17)31416-9
DOI: 10.1016/j.neucom.2017.04.077
Reference: NEUCOM 18806

To appear in: *Neurocomputing*

Received date: 18 April 2016
Revised date: 22 March 2017
Accepted date: 3 April 2017

Please cite this article as: Huaping Liu, Fengxue Li, Xinying Xu, Fuchun Sun, Multi-Modal Local Receptive Field Extreme Learning Machine for Object Recognition, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.04.077

Multi-Modal Local Receptive Field Extreme Learning Machine for Object Recognition

Huaping Liu[a,*], Fengxue Li[a], Xinying Xu[a], Fuchun Sun[a]

[a]*Department of Computer Science and Technology, Tsinghua University, State Key Lab. of Intelligent Technology and Systems, TNLIST, Beijing, P.R. China*

Abstract

Learning rich representations efficiently plays an important role in the multi-modal recognition task, which is crucial to achieving high generalization performance. To address this problem, in this paper, we propose an effective Multi-Modal Local Receptive Field Extreme Learning Machine (MM-LRF-ELM) structure, while maintaining ELM's advantages of training efficiency. In this structure, LRF-ELM is firstly conducted for feature extraction for each modality separately. And then, the shared layer is developed by combining these features from each modality. Finally, the Extreme Learning Machine (ELM) is used as supervised feature classifier for the final decision. Experimental validation on Washington RGB-D Object Dataset illustrates that the proposed multiple modality fusion method achieves better recognition performance.

Keywords: Representation learning, multi-modal, local receptive field, extreme learning machine.

1. Introduction

Object recognition is a challenging task in computer vision and important for making robots useful in home environments. With the recent advent of depth cameras, an increasing amount of visual data not only contains color but also depth measurements. Compared to RGB data, which provides information about appearance and texture, depth data contains additional

[*] Corresponding author
Email address: hpliu@tsinghua.edu.cn (Huaping Liu)

information about object shape and it is invariant to lighting or color variations [1].

In recent years, various approaches that have been proposed for RGB-D object recognition: methods with hand-crafted features [2, 3, 4], and methods with learned feature [5, 6, 7, 8, 9, 10]. Moreover, the classical neural network structure, like convolutional neural network networks (CNNs), is also applied to the object recognition field [24, 25, 26] and it have recently been shown to be remarkably successful for recognition on RGB images [23].

Though traditional gradient-based learning algorithms (like BP Neural network) [11] have been widely used in the training of multilayer feedforward neural networks [21, 22], these gradient-based learning algorithms are still relatively slow in learning and easily get stuck in local minima [13]. Furthermore, the activation functions used in these gradient-based tuning methods need to be differentiable.

In order to overcome the drawbacks of gradient-based methods, Huang et al. proposed an efficient training algorithm for the single-hidden layer feedforward neural network (SLFN) called Extreme Learning Machine (ELM) [12, 14]. It increases the learning speed by means of randomly generating input weights and hidden biases, and the output weights are determined by using Moore-Penrose (MP) generalized inverse. Compared with the traditional gradient-based learning algorithms, ELM not only learns much faster with higher generalization performance [27, 30] but also avoids many difficulties faced by gradient-based learning methods such as stopping criteria, learning rate, learning epochs, and local minima. What's more, more and more deep ELM learning algorithms has been proposed [33, 34] to capture relevant higher-level abstractions. However, ELM with local connections has not attracted much research attention yet. Ref. [15] has proved that the application of the local receptive fields based ELM (LRF-ELM) has better performance than conventional deep learning solutions [16, 31, 32] in image processing and speech recognition.

However, the aforementioned works do not refer to the multi-modal problem [28, 29]. Thus, in this paper, we extend the LRF-ELM and propose a Multi-Modal LRF-ELM (MM-LRF-ELM) framework. The proposed MM-LRF-ELM is applied to multi-modal learning task, while maintaining its advantages of training efficiency. The contributions of this work are summarized as follows:

1. We propose an architecture: multi-modal LRF-ELM framework, to

construct the nonlinear representation from different aspects of information sources. The important merit of such a method is that the training time is greatly shortened and the testing efficiency is highly improved.

2. We evaluate our multimodal network architecture on the Washington RGB-D Object Dataset [4]. The obtained results show that the proposed fusion method obtains rather promising results.

The remainder of this paper is organized as follows: Section 2 introduces the related works, including the fundamental concepts and theories of ELM; Section 3 describes the proposed MM-LRF-ELM framework; Section 4 compares the performance of MM-LRF-ELM with single modality and other methods; while Section 5 concludes this paper.

2. Brief review for ELM

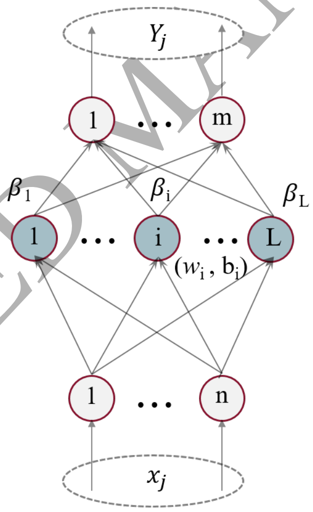


Figure 1: The model of basic ELM

ELM was proposed in Huang, et al [12]. Suppose we are training SLFNs with K hidden neurons and activation function g(x) to learn N distinct samples $\{\mathbf{X}, \mathbf{T}\} = \{\mathbf{X}_j, \mathbf{t}_j\}_{j=1}^{N}$,where $\mathbf{x}_j \in \mathbf{R}^n$ and $\mathbf{t}_j \in \mathbf{R}^m$. In ELM, the input weights and hidden biases are randomly generated instead of tuned. By doing so, the nonlinear system has been converted to a linear system

$$\mathbf{Y}_j = \sum_{i=1}^{L} \beta_i g_i(\mathbf{x}_j) = \sum_{i=1}^{L} \beta_i g(\mathbf{w}_i^T \mathbf{x}_j + \mathbf{b}_i) = t_j, j = 1, 2, ...N \qquad (1)$$

where $\mathbf{Y}_j \in \mathbf{R}^m$ is the output vector of the j-th training sample, $\mathbf{W}_i \in \mathbf{R}^n$ is the input weight vector connecting the input nodes to the i-th hidden node,b_i denotes the bias of the i-th hidden neuron;$\beta_i = (\beta_{i1}, \beta_{i2}, ..., \beta_{im})^T$ denotes the weight vector connecting the i-th hidden neuron and output neurons; $g(\cdot)$ denotes hidden nodes nonlinear piecewise continuous activation functions. The above N equations can be written compactly as:

$$\mathbf{H}\beta = \mathbf{T} \qquad (2)$$

where the matrix T is target matrix,

$$\mathbf{H} = \begin{bmatrix} g(\mathbf{w}_1^T \mathbf{x}_1 + \mathbf{b}_1) & \cdots g(\mathbf{w}_L^T \mathbf{x}_1 + \mathbf{b}_L) \\ \vdots & \cdots \quad \vdots \\ g(\mathbf{w}_1^T \mathbf{x}_N + \mathbf{b}_1) & \cdots g(\mathbf{w}_L^T \mathbf{x}_N + \mathbf{b}_L) \end{bmatrix} \qquad (3)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}, \ \mathbf{T} = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix} \qquad (4)$$

Thus, the determination of the output weights (linking the hidden layer to the output layer) is as simple as finding the least-square solution to the given linear system. The minimum norm least-square (LS) solution to the linear system (1) is

$$\hat{\beta} = \mathbf{H}^{\dagger}\mathbf{T} \qquad (5)$$

where $H^{\dagger}$ is the MP generalized inverse of matrix H. As analyzed by Huang, et al., ELM using such MP inverse method tends to obtain good generalization performance with dramatically increased learning speed.

3. Multi-Modal LRF-ELM

3.1. Model Architecture

Our architecture, which is depicted in Fig.2, employs the LRF-ELM as the learning unit to learn shallow and deep information. 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.2, we perform feature learning to have representations of each modality (RGB and Depth) before they are mixed. Each modality is given to a single LRF-ELM net layer which provides useful translational invariance of low-level features such as edges and allows parts of an object to be deformable to some extent.

Mathematically, the output of each modality can be separately calculated. where $\mathbf{H}_1^c, \mathbf{H}_2^d \in N \times K \cdot (d-r+1)^2$,the parameter N is the input samples, K is the number of feature maps , d is the input size and r is the size of the receptive field.$\mathbf{H}_1^c, \mathbf{H}_2^d$ are the pooling layer feature matrixes representing non-linear representations extracted from features of each modality, where c denotes the LRF-ELM I, which extracts the feature of the RGB image and d denotes the LRF-ELM II, which extracts the feature of the Depth image. In our work, each LRF-ELM net layer has the same parameters.

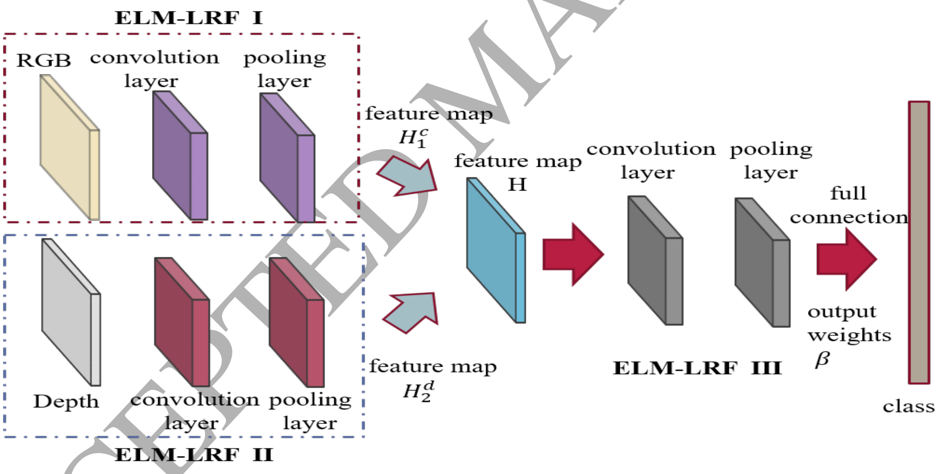


Figure 2: The proposed Multi-Modal architecture

A single LRF-ELM net layer extracts low level features from RGB and depth images respectively. Both representations are given as input to another LRF-ELM layer, the combination process is as follows:

$$\mathbf{H} = \left[\mathbf{H}_1^c; \mathbf{H}_2^d\right]^T \tag{6}$$

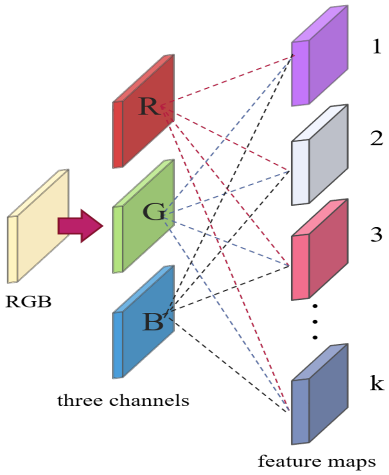


Figure 3: The convolution process of RGB image

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

3.2. Unsupervised Feature Representation

In this work, we adopt the local receptive fields based on ELM (LRF-ELM) to extract the features. In LRF-ELM, the links between input and hidden layer nodes are sparse and bounded by corresponding receptive fields, which are be sampled from any continuous probability distribution [15]. Fig.4 illustrates that the process of learning representation from the features of each modality.

The LRF-ELM consists of two basic operations:

(1) Generate the initial weight matrix $\hat{\mathbf{A}}^{c}_{init}$, $\hat{\mathbf{A}}^{d}_{init}$ randomly.With the input size $d \times d$ and the receptive field $r \times r$, the size of the feature map should be $(d-r+1) \times (d-r+1)$.

$$\begin{aligned} &\hat{\mathbf{a}}^{c}_{k},\, \hat{\mathbf{a}}^{d}_{k} \in \mathbf{R}^{r^2} \\ &\hat{\mathbf{A}}^{c}_{init}, \hat{\mathbf{A}}^{d}_{init} \in \mathbf{R}^{r^2 \times k}, k = 1, 2, 3...K \end{aligned} \tag{7}$$

then, orthogonalize the initial weight matrix $\hat{\mathbf{A}}^{c}_{init}$,$\hat{\mathbf{A}}^{d}_{init}$,using singular value decomposition (SVD) method.

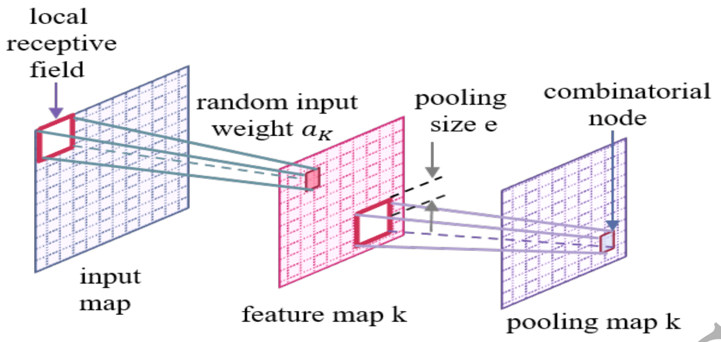


Figure 4: The LRF-ELM architecture

The input weight to the k-th feature map is $\mathbf{a}_k^c$,$\mathbf{a}_k^d \in \mathbf{R}^{r\times r}$,which corresponds to $\hat{\mathbf{a}}_k^c$,$\hat{\mathbf{a}}_k^d \in \mathbf{R}^{r^2}$,column-wisely.The convolutional node (i,j) in the k-th feature map $\mathbf{C}_{i,j,k}$ is calculated as:

$$\begin{cases} \mathbf{C}^{(c)}{}_{i,j,k}(x) = \sum\limits_{m=1}^{r}\sum\limits_{n=1}^{r} \mathbf{x}^{c}{}_{i+m-1,j+n-1}\cdot \mathbf{a}^{c}{}_{m,n,k} \\ \mathbf{C}^{(d)}{}_{i,j,k}(x) = \sum\limits_{m=1}^{r}\sum\limits_{n=1}^{r} \mathbf{x}^{d}{}_{i+m-1,j+n-1}\cdot \mathbf{a}^{d}{}_{m,n,k} \end{cases}$$
$$i,j=1,...,(d-r+1) \tag{8}$$

(2) Formulate the combinatorial node. Pooling size e is the distance between the center and the edge of the pooling area. And the pooling map is of the same size with the feature map.$\mathbf{C}_{i,j,k}^c$,$\mathbf{C}_{i,j,k}^d$ and $\mathbf{h}_{p,q,k}^c$,$\mathbf{h}_{p,q,k}^d$, respectively denote node(i,j) in the k-th feature map and combinatorial node (p, q) in the k-th pooling map:

$$\begin{cases} \mathbf{h}^{(c)}{}_{p,q,k} = \sqrt{\sum\limits_{i=p-e}^{p+e}\sum\limits_{j=q-e}^{q+e} \mathbf{C}^{(c)2}{}_{i,j,k}} \\ \mathbf{h}^{(d)}{}_{p,q,k} = \sqrt{\sum\limits_{i=p-e}^{p+e}\sum\limits_{j=q-e}^{q+e} \mathbf{C}^{(d)2}{}_{i,j,k}} \end{cases}$$
$$p,q=1,...,(d-r+1),$$
$$if(i,j)\ is\ out\ of\ bound: \mathbf{C}^{(c)}{}_{i,j,k}, \mathbf{C}^{(d)}{}_{i,j,k}=0 \tag{9}$$

3.3. Supervised Feature Classification

Through the unsupervised feature representation, we can get a part of the parameters need to be learned. The pooling layer in ELM-LRF III is

in full connection with the output layer.Simply concatenating the values of all combinatorial nodes into a row vector and putting the rows of N' input samples together. Here to improve generalization performance and make the solution more robust, we can add a regularization term [20]. Then obtain the combinatorial layer matrix $\mathbf{H}' \in R^{N' \times K \cdot (d'-r+1)^2}$:
1.if $N' \leq K \cdot (d' - r + 1)^2$

$$\beta = \mathbf{H}'^T (\frac{1}{C} + \mathbf{H}'\mathbf{H}'^T)^{-1}\mathbf{T} \tag{10}$$

2.if $N' \geq K \cdot (d' - r + 1)^2$

$$\beta = (\frac{1}{C} + \mathbf{H}'^T\mathbf{H}')^{-1}\mathbf{H}'^T\mathbf{T} \tag{11}$$

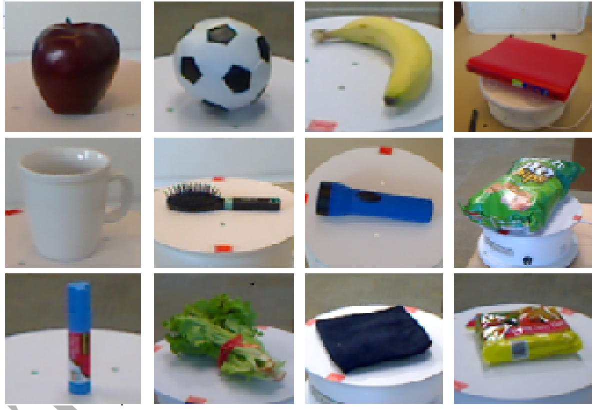


Figure 5: Some objects from the RGB-D Object Dataset

4. Experimental results

4.1. Data Set

The Washington RGB-D Object Dataset consists of 41,877 RGB-D images containing household objects organized into 51 different classes and a total of 300 instances of these classes which are captured under three different viewpoint angles. For the evaluation, every 5th frame is subsampled.

Our experiments focused on category recognition and instance recognition. After subsampling every 5th frame from the videos, there were some

Table 1: Category recognition accuracy for different approaches

Method	**RGB**	**Depth**	**RGB-D**
Nonlinear SVM [4]	74.5 ± 3.1	64.7 ± 2.2	83.9 ± 3.5
HKDES [17]	76.1 ± 2.2	75.7 ± 2.6	84.1 ± 2.2
Kernel Desc [2]	77.7 ± 1.9	78.8 ± 2.7	86.2 ± 2.1
CKM Desc [5]	N/A	N/A	86.4 ± 2.3
CNN-RNN [8]	80.8 ± 4.2	78.9 ± 3.8	86.8 ± 3.3
Upgraded HMP [7]	82.4 ± 3.1	81.2 ± 2.3	87.5 ± 2.9
CaRFs [18]	N/A	N/A	88.1 ± 2.4
CNN Features [19]	83.1 ± 2.0	N/A	89.4 ± 1.3
HHA[35]	84.1 ± 2.7	83.0 ± 2.7	91.0 ± 1.9
MM-LRF-ELM	$\mathbf{84.3 \pm 3.2}$	$\mathbf{82.9 \pm 2.5}$	$\mathbf{89.6 \pm 2.5}$
Shallow Combination	N/A	N/A	74.3 ± 2.6

Table 2: The training time for the category recognition

Method	**RGB**	**Depth**	**RGB-D**
MM-LRF-ELM	573.44s	435.36s	715.66s
CNN	8716.28s	6138.76s	12023.09s

34000 images for training and 6900 images for testing. Before the images given to the LRF-ELM, they are resized to be $d = 128$. Parameters that need in LRF-ELM: the size of the receptive field is 5×5. the pooling size $e = 3$, the value of the balance parameter $C = 0.01$.

4.2. Category Recognition

We compare our fusion network with other approaches reported for the RGB-D dataset.We adopted the same setup as [4] and ran the 10 random splits provided, resulting in 51 test objects. We also use the Shallow Combination method, which performs multi-modal fusion by combining RGB and Depth features as a concatenated vector that acts as the input of the framework. Results are recognition accuracy in percent. Our MM-LRF-ELM outperforms the previous approaches.

Tab.1 summarizes that compared with other methods, our multi-modal method can achieve better performance. The results show that extracting features independently from RGB channel and Depth channel and then combing them at the later stage may result in better performance. The main reason is that the features which are extracted from different channels at early

Table 3: Instance recognition accuracy for different approaches

Method	RGB	Depth	RGB-D
Nonlinear SVM [4]	60.7	46.2	74.8
HKDES [17]	79.3	46.8	82.4
Kernel Desc [2]	90.8	54.3	91.2
CKM Desc [5]	82.9	N/A	90.4
Upgraded HMP [7]	92.1	51.7	92.8
CNN Features [19]	92.0	N/A	94.1
MM-LRF-ELM	**91.0**	**50.9**	**92.5**
Shallow Combination	N/A	N/A	84.5

Table 4: The training time for the instance recognition

Method	RGB	Depth	RGB-D
MM-LRF-ELM	455.65s	471.33s	685.66s
CNN	8247.26s	8201.15s	10285.4s

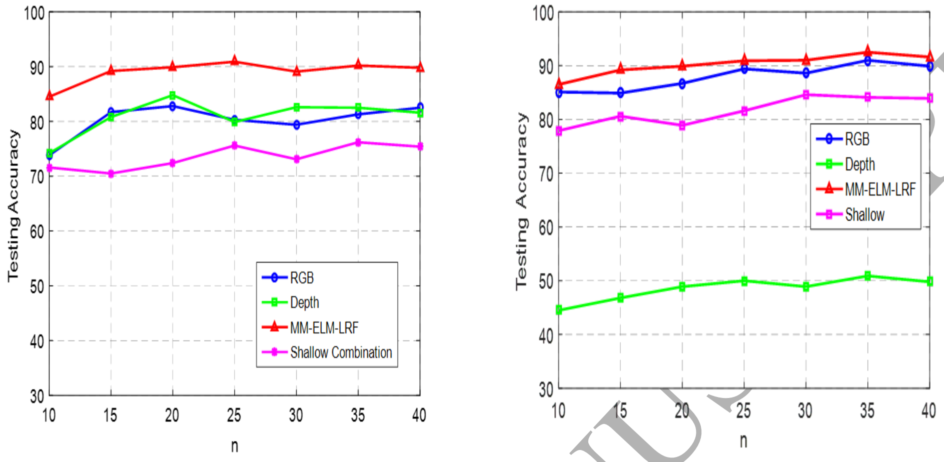
stages may result in the features being more independent.

Table.1 shows that our model is comparable to the state-of-the-art. In addition, to compare the time costs, we use the MatConv Toolbox in MATLAB to implement the CNN-based method. The training time costs for the single modal case and multi-modal fusion are listed in Table.2, which shows that the proposed method significantly reduces the training time costs.

4.3. Instance Recognition

Following the experimental setting suggested by [4], we train models on the video sequences of each object where the viewing angles are 30° and 60° with the horizon and test them on the 45° video sequence. As can be seen in Table.3, For object recognition on the instance level, the proposed method also achieves comparable results. In addition, the time costs listed in Table 4 shows that advantages of the proposed method.

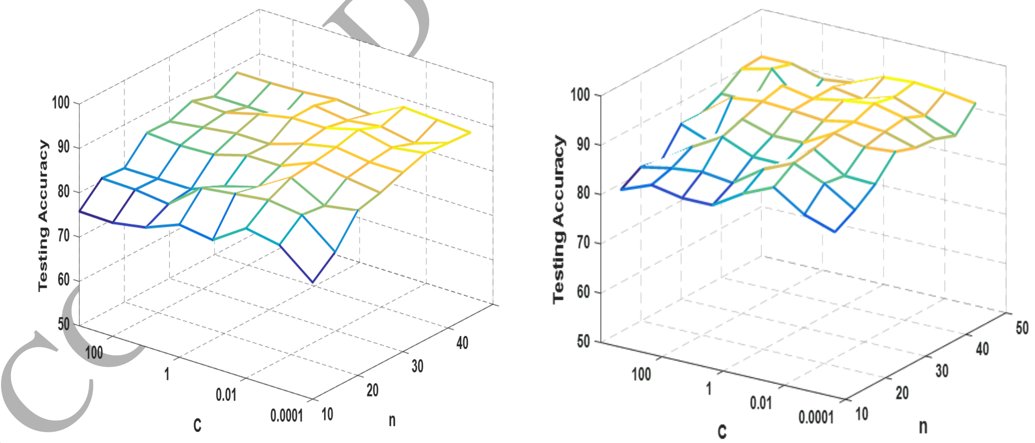
Fig.6 shows that compared with single modality networks, our MM-LRF-ELM is able to learn rich representations efficiently which outperforms single modal ones for recognition. The Shallow Combination method in category



Category recognition

Instance recognition

Figure 6: The testing accuracies of different methods versus the number of feature maps



Category recognition

Instance recognition

Figure 7: Testing accuracy of MM-LRF-ELM in terms of n and C

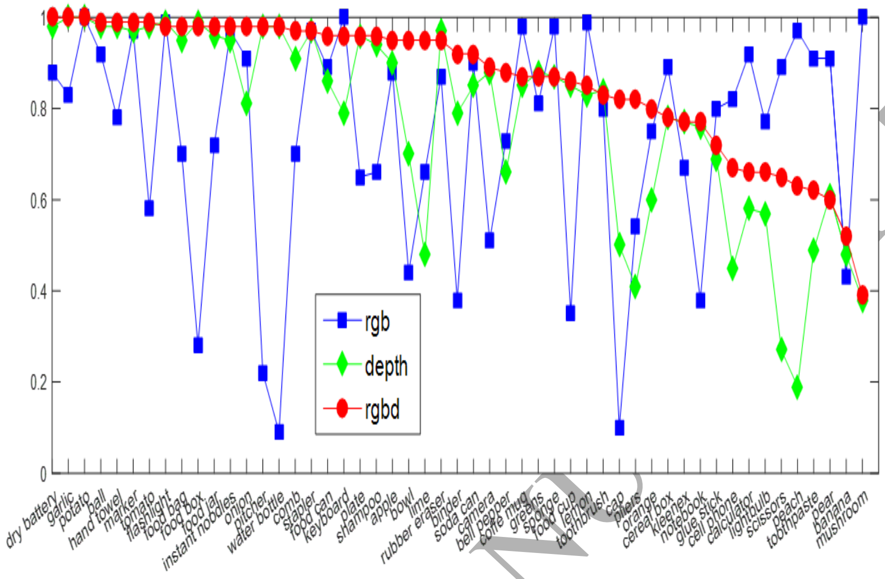


Figure 8: Per-class testing accuracy of our model compared with single modal

recognition shows poor classification performance because the great difference between each modality. We also notice that depth features are much worse than image features in the context of instance recognition. It is not very surprising since the different instances have very similar shape in the same category.

To analyze the roles of the parameters, we perform the sensitivity analysis. The most important two parameters in the proposed MM-LRF-ELM include the balance parameter C and the number of feature map n. Therefore, we vary the value of C within the set$\{10^{-4}, 10^{-3}, 10^{-2}, ...10^{2}, 10^{3}\}$ and the value of n within the set $\{10, 15, 20, 25, ...45\}$to analyze the performance variations. The results are shown in Fig.7. With the single-modal and another method, the training of MM-LRF-ELM is much faster and achieves higher learning accuracy.

4.4. Failure Analysis

From Fig.8, we can see that our model MM-LRF-ELM greatly improves the accuracy compared with the single modality and the worst class recall

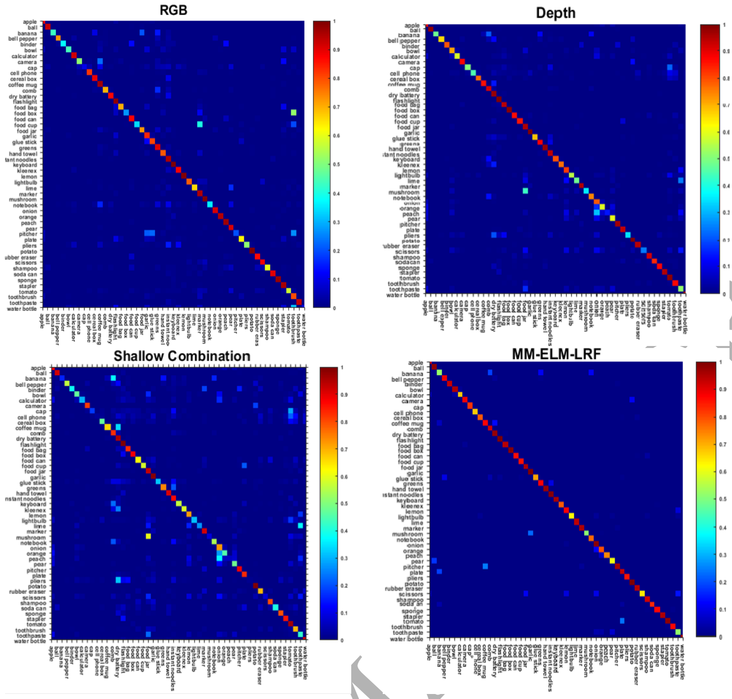


Figure 9: Confusion matrix of the category recognition on RGB-D Object Dataset. The vertical axis shows the true labels and the horizontal axis shows the predicted labels. Please refer to the electronic version for an enlarged illustration.

belongs to mushrooms which are very similar in appearance to garlic.We also analyze the overall performance of our model depicted in Fig.10. Our model shows only one outlier which belongs to the mushroom. The average classification accuracy and the lowest classification accuracy are much higher than that of RGB and Depth and the Shallow Combination method shows the poor overall classification performance.

Fig.9 shows the confusion matrix across all 51 classes. Most model confusions are very reasonable showing that the MM-LRF-ELM method can give high-quality features compared with the single modality and the Shallow Combination method.

Fig.11 shows 6 pairs of confused classes. For the category recognition, mushrooms labeled as garlic, pitchers classified as coffee mug due to shape and color similarity, toothpastes classified as pears at certain angles. For the instance recognition, different instances in the same category may have the similar shape ,for example, the second class of bananas assigned to the wrong

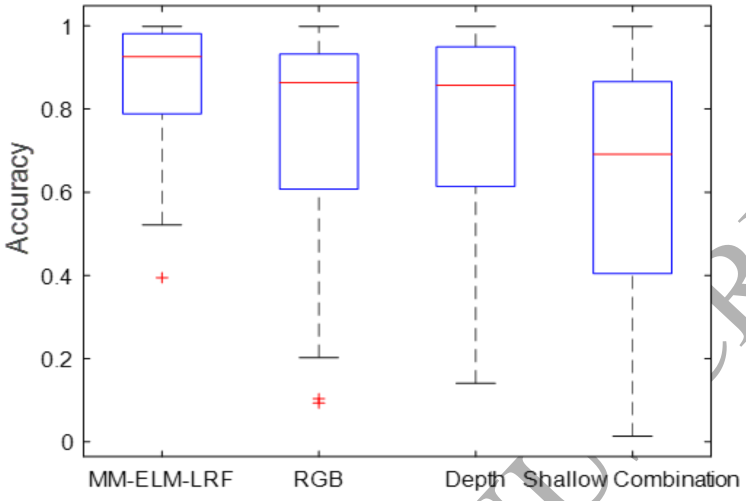


Figure 10: The comparison of the overall classification performance

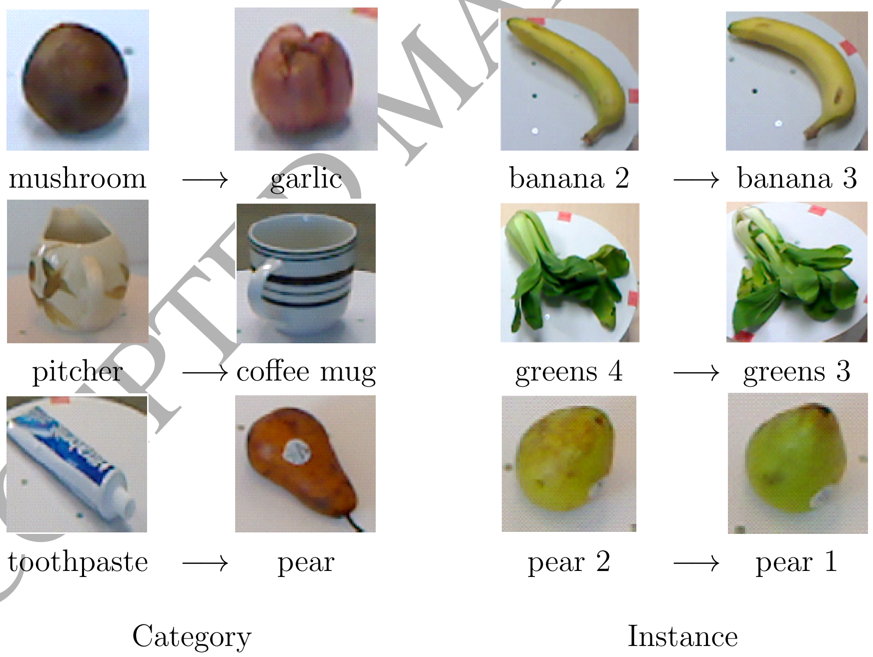


Figure 11: Examples of some confused classes

class three.

5. Conclusions

In this paper, we have proposed a novel multi-modal training scheme MM-LRF-ELM, in which information of each modality has been learned and combined in an effective way without iterative fine-tuning. In this structure, MM-LRF-ELM takes full advantage of the LRF-ELM to learn the high-level representation of the multi-modal data. Thus,the proposed method could obtain more robust and better performance.

6. Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants U1613212, 61673238, 91420302, and 61327809, in part by the National High-Tech Research and Development Plan under Grant 2015AA042306.

7. References

References

[1] M. Quigley, S. Batra, S. Gould, E. Klingbeil, Q. Le, A. Wellman, and A.Y.Ng. "High-accuracy 3D sensing for mobile manipulation:improving object detection and door opening." *in Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, pp.2816-2822, 2009.

[2] L. Bo, X. Ren and D. Fox. "Depth kernel descriptors for object recognition." *in Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp.821-826, 2011.

[3] B. Browatzki, J. Fischer, B. Graf, H. Bulthoff, and C. Wallraven. "Going into depth: Evaluating 2d and 3d cues for object classification on a new, large-scale object dataset." *in Proc. of the Computer Vision Workshops (ICCV Workshops)*, pp.1189-1195, 2011.

[4] K. Lai, L. Bo, X. Ren, and D. Fox. "A large-scale hierarchical multi-view rgb-d object dataset." *in Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, pp. 1817-1824, 2011.

[5] M. Blum, J. T.Springenberg, J. Wuelfing, and M. Riedmiller. "A learned feature descriptor for object recognition in rgb-d data." *in Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, pp.1298-1303, 2012.

[6] L. Bo, X. Ren, and D. Fox. "Hierarchical matching pursuit for image classification: Architecture and fast algorithms." *in Proc. of the Neural Information Processing Systems (NIPS)*, pp.2115-2123, 2011.

[7] L. Bo, X, Ren, and D. Fox. "Unsupervised feature learning for rgb-d based object recognition." *in Proc. of the Int. Symposium on Experimental Robotics (ISER)*, pp.387-402, 2012.

[8] R. Socher, B. Huval, B. Bhat, C. D. Manning and A. Y. Ng. "Convolutional -recursive deep learning for 3d object classification." *in Proc. of the Neural Information Processing Systems (NIPS)*, pp.665-673, 2012.

[9] J. Wang, J, Yang, K, Yu, F. Lv, T. Huang, and Y. Gong. "Locality constrained linear coding for image classification." *in Proc. of the Int. Conf.on Computer Vision and Pattern Recognition (CVPR)*, pp.3360-3367, 2010.

[10] K. Yu, Y. Lin, and J. Lafferty. "Learning image representations from the pixel level via hierarchical sparse coding." *in Proc. of the Computer Vision and Pattern Recognition (CVPR)*, pp.1713-1720, 2011.

[11] D. E. Rumelhart, J. L.Mcclelland. "Parallel distributed procesing." *Encyclopedia of Database Systems*, pp.45-76, 1986.

[12] G. Huang, Q. Zhu, C. Siew. "Extreme learning machine: a new learning scheme of feedforward neural networks." *in Proc. of the International Joint Conference Neural Networks (IJCNN)*, pp.985-990, 2004.

[13] G. Huang, Q. Zhu, C. Siew. "Extreme learning machine: Theory and applications." *Neurocomputing*, pp.489-501, 2006.

[14] G. Huang, H. Zhou, X. Ding, R. Zhang. "Extreme Learning Machine for Regression and Multi-class Classification." *in Proc. of the IEEE Systems Man & Cybernetics Society*, pp.513-529, 2012.

[15] G. Huang, Z. Bai, L. L. C. Kasun, M. V. Chi. "Local receptive fields based extreme learning machine." *in Proc. of the IEEE Computational Intelligence Magazine*, pp.18-29, 2015.

[16] K. Hornik. "Approximation capabilities of multilayer feedforwardnetworks." *Neural Networks*, pp.251-257, 1991.

[17] L. Bo, K. Lai, X. Ren, and D. Fox. "Object recognition with hierarchical kernel descriptors." *in Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.1729-1736, 2011.

[18] U. Asif, M. Bennamoun, and F. Sohel. "Efficient rgb-d object categorization using cascaded ensembles of randomized decision trees." *in Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, pp.1295-1302, 2015.

[19] M. Schwarz, H. Schulz, and S. Behnke. "RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features." *in Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, pp.1329-1335, 2015.

[20] S. Ding, N. Zhang, X. Xu, L. Guo, J. Zhang. "Deep extreme learning machine and its application in EEG classification." *Mathematical Problems in Engineering*, pp.1-11, 2014.

[21] G. Huang, H. A. Babri. "Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions." *IEEE Trans. Neural Networks*, pp.224-229, 1998.

[22] M. Leshno, V. Y. Lin, A. Pinkus, S. Schocken. "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function." *Neural Networks*, pp.861-867, 1993.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks." *in Proc. of the Neural Information Processing Systems (NIPS)*, pp.1097-1105, 2012.

[24] B. Hariharan, P. Arbelez, R. Girshick, and J. Malik. "Simultaneous detection and segmentation." *in Proc. of the European Conference on Computer Vison (ECCV)*, pp.297-312, 2014.

[25] K. Simonyan and A. Zisserman. “Two-stream convolutional networks for action recognition in videos.” *in Proc. of the Neural Information Processing Systems (NIPS)*, pp.568-576, 2014.

[26] N. Srivastava and R. R. Salakhutdinov. “Multimodal learning with deep boltzmann machines.” *in Proc. of the Neural Information Processing Systems (NIPS)*, pp.1967-2006, 2012.

[27] M. B. Li, G. Huang, P. Saratchandran, N. Sundararajan. “Fully complex extreme learning machine.” *Neurocomputing*, pp.306-314, 2005.

[28] J. Tang, Z. Li, M. Wang, and R. Zhao. “Neighborhood discriminant hashing for large-scale image retrieval.” *IEEE Transactions on Image Processing*, pp.2827-2840, 2015.

[29] W. Lu, J. Li, W. Guo, H. Zhang, and J. Guo. “Web multimedia object classification using cross-domain correlation knowledge.” *IEEE Transactions on Multimedia*, pp.1920-1929, 2013.

[30] X. Wang, M. Han. “Multivariate time series prediction based on multiple kernel extreme learning machine.” *in Proc. of the International Joint Conference Neural Networks (IJCNN)*, pp.198-201, 2015.

[31] W. Huang, H. Hong, G. Song and K.Xie. “Deep process neural network for temporal deep learning.” *in Proc. of the International Joint Conference Neural Networks (IJCNN)*, pp.465-472, 2014.

[32] K. H. Cho, T. Raiko, A. Llin. “Gaussian-bernoulli deep bolzmann machine.” *in Proc. of the International Joint Conference Neural Networks (IJCNN)*, pp.1-7, 2013.

[33] L. L. C. Kasun, H. Zhou, G. Huang, and C. M. Vong, “Representational Learning with Extreme Learning Machine for Big Data,” *IEEE Intelligent Systems*, pp.31-34, 2013.

[34] Y. Yang and Q. M. J. Wu, “Mutilayer Extreme Learning Machine with Subnetwork Nodes for Representation Learning.” *IEEE Transactions on Cybernetics*, pp.1-14, 2015.

[35] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller and W. Burgard, “Multimodal deep learning for robust RGB-D object recognition,” *in Proc. of the Intelligent Robots and Systems (IROS)*, pp.681-687.

Huaping Liu is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His current research interests include robot perception and learning. He serves as an Associate Editor of some journals, including the IEEE ROBOTICS AND AUTOMATION LETTERS, Neurocomputing, the International Journal of Control, Automation and Systems.

Fengxue Li graduated from School of Information Sciences, Taiyuan University of Technology from 2013. She is now graduate student and her interests are machine learning and its applications.

Xinying Xu is an associate professor in School of Information Sciences, Taiyuan University of Technology. His research interests are machine learning and its applications.

Fuchun Sun is currently a Full Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His current research interests include intelligent control and robotics. He was a recipient of the National Science Fund for Distinguished Young Scholars. He serves as an Associate Editor of a series of international journals, including the IEEE TRANSACTIONS ON FUZZY SYSTEMS, Mechatronics, Robotics, and Autonomous Systems.