Multimed Tools Appl (2017) 76:4273–4290
DOI 10.1007/s11042-016-3375-5

Modality-specific and hierarchical feature learning for RGB-D hand-held object recognition

Xiong Lv[1] · Xinda Liu[2] · Xiangyang Li[1] · Xue Li[1,3] · Shuqiang Jiang[1] · Zhiqiang He[4]

Received: 3 December 2015 / Revised: 9 February 2016 / Accepted: 15 February 2016 /
Published online: 4 March 2016

Abstract Hand-held object recognition is an important research topic in image understanding and plays an essential role in human-machine interaction. With the easily available RGB-D devices, the depth information greatly promotes the performance of object segmentation and provides additional channel information. While how to extract a representative and discriminating feature from object region and efficiently take advantage of the depth information plays an important role in improving hand-held object recognition accuracy and eventual human-machine interaction experience. In this paper, we focus on a special but important area called RGB-D hand-held object recognition and propose a hierarchical

✉ Zhiqiang He
lirong2@lenovo.com

Xiong Lv
XiongLv@vipl.ict.ac.com

Xinda Liu
yunonglxd@163.com

Xiangyang Li
XiangyangLi@vipl.ict.ac.com

Shuqiang Jiang
shuqiang.jiang@vipl.ict.ac.cn

Xue Li
XueLi@vipl.ict.ac.com

[1] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

[2] School of Mathematics and Computer Science, Ningxia University, Ningxia 750021, China

[3] College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong Province, China

[4] Lenovo Corporate Research, Beijing 100085, China

feature learning framework for this task. First, our framework learns modality-specific features from RGB and depth images using CNN architectures with different network depth and learning strategies. Secondly a high-level feature learning network is implemented for a comprehensive feature representation. Different with previous works on feature learning and representation, the hierarchical learning method can sufficiently dig out the characteristics of different modal information and efficiently fuse them in a unified framework. The experimental results on HOD dataset illustrate the effectiveness of our proposed method.

Keywords Feature learning · RGB-D object recogntion · Multiple modalities

1 Introduction

Objects are playing an important role in human-machine interaction, which helps machine to better understand the environment. Therefore object recognition [2, 3, 6] has been a significant research field in computer vision. It can assist image content understanding, scene modeling, multimedia retrieval and so on. At the same time, during human-machine interaction in AI system, the object held by user is not a negligible factor in understanding the user's intention and requirements. Imagining that when you ask an AI system "I want some thing like this" with a book named "Harry Potter" on your hand, the system should combine the object you hold and your question together to understand your requirement, then finds a similar one and shows it to you. Therefore, there rises a special but important area called hand-held object recognition, which focuses on recognizing the object held on user's hand. It can not only help machine to "see" the object that may be correlated to user's intention but also make a more specific inference or reasonable reaction about user's requirements.

Research on hand-held object recognition [1, 12, 14, 15, 18] can be divided into two categories: one is the first-person interface [1, 18], in which the hand-held images are captured from the first-person point of view; another is second-person interface [12, 14, 15], which uses a camera located in the robot or system for the user to interact with. In first-person view, the image is always captured from a smartphone and only has RGB information. In second-person view, the captured image often contains object, person and background, the object held in hand may only occupy a small region of the image. As RGB-D devices (e.g. Kinect, RealSense) are more and more common and inexpensive, it's convenient to capture the RGB and depth information from real scene. Some RGB-D devices can also provide skeletal information of the user, which helps the system to know user's relation with environment, especially the hand-held object. The advantages of depth information include: 1) providing depth information of each pixel in RGB image, which does not exist in traditional RGB image; 2) depth information naturally contains the object regions (the region area in same depth level can be a potential object or object set); 3) depth information of an object region can represent shape information of object surface. Because second-person interface is widely used in human-machine interaction, we focus on this task in this paper.

Many works [12, 14, 15] on RGB-D images take advantage of RGB-D device to segment the target object region using depth and skeletal information. Liu et al. [12] first build a point cloud based on RGB and depth information, then extract ESF [21], C^3-HALC [9] and GRSD [16] features, finally implement multiple kernel learning [4] (MKL) for feature fusion and SVM for classification. Lv et al. [14] add deep learned features in the fusion model and concatenate the deep learned features on depth and RGB images in the first stage, then fuse them with hand-crafted features using MKL. A common pipeline of RGB-D

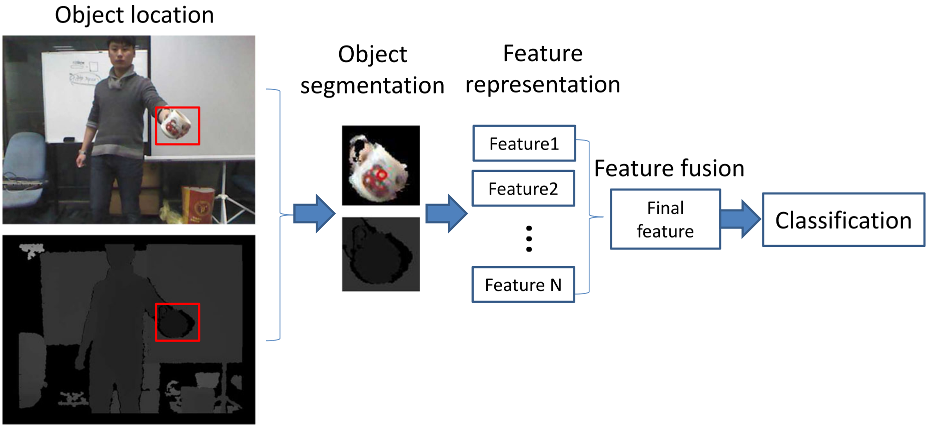


Fig. 1 RGB-D hand-held object recognition pipeline

hand-held object recognition is shown in Fig. 1. These works combine the depth and RGB information either on point cloud level or feature level. While hand-crafted features are various and each of them describes one or several aspects of the image (for example ESF mainly represents the shape information, C^3-HALC represents color information), which makes it difficult to select a suitable combination of features to have a comprehensive and discriminative representation of the dataset. Meanwhile the feature designing procedure is timeconsuming and has less scalability. With the explosion of deep learning (i.e.Convolutional neural networks (CNN)), it tremendously alleviates the complexity in designing a representative feature by learning feature from the image dataset itself. In [14], a CNN model is trained on the ImageNet2012 dataset, then the model is directly used to extract features from RGB and depth image in HOD dataset. Because the training dataset is not HOD, which will reduce the representative ability in target dataset, therefore Lv et al. [14] add hand-crafted features in [12] and use MKL to fuse them and use SVM for classification.

Motivated by previous works on RGB-D hand-held object recognition, we want to find a way that can learn features on RGB and depth images from the RGB-D hand-held object dataset, and automatically learn a high-level feature using two modal features. Therefore, in this paper we propose a modality-specific and hierarchical feature learning method for RGB-D hand-held object recognition. The method implements two networks in different depths for RGB and depth feature learning respectively, and a third network for high-level feature learning. The main contributions in this paper are as follows:

- We propose a new feature learning and representation framework for RGB-D hand-held object recognition based on CNN in a unified framework. It can learn distinctive and representative features on RGB and depth information via different network architectures and learning schemes. It can sufficiently exploit the complementary factor between different features, and learn a more comprehensive and high-level feature via hierarchical networks.
- Modality-specific CNN architectures and modality-correlated learning schemes (msCNN) are proposed for RGB and depth feature learning. It uses external dataset to pretrain the networks to obtain basic and shared features. Then we finetune the two networks in different ways to learn modality-correlated features. The feature learning

method is proved to have an improved representative ability than directly extracting features using ready-made networks.

This paper is organized as follows. The second section shows some related works. The Sections 3, 4, 5 sections detail the framework, modality-specific feature learning and high-level feature learning respectively. Section 6 introduces the experimental results. The last section summarizes the proposed method and discusses the future work.

2 Related works

2.1 3D convolutional neural network

There are several works on CNN which take two or more input images as input instead of one RGB image. Ji et al. [8] show a work on action recognition, which combines multiple images as a whole input for CNN. It changes the first convolutional layer of normal CNN by using convolutional kernels to convolute multiple input images as one image to generate the feature map. Gupta et al. [7] convert the depth image into a three dimensional image called HHA. Firstly, they encode depth image with three channels at each pixel: horizontal disparity, height above ground, and the angle of the pixel's local surface normal. Secondly, they use two same networks and train them on both HHA and RGB images respectively. Finally, they concatenate the two features and feed them into a SVM classifier. Ji et al. [8] focus on action recognition, they use multiple RGB images as input for CNN, the RGB images have some common patterns like the color filter and contour. While our work focuses on RGB-D images, RGB and depth images have different modal information, directly using a filter to convolute them will ignore the difference and make it hard to find the characteristic of each modality. Although Gupta et al. [7] take use of depth information by converting the depth image into a new form of image called HHA, which makes the generated image have similar pattern with RGB image. While this is very complex and converting will change the original information of the depth image. Our method takes raw depth images and RGB images as input, which will not reduce or change the input information.

2.2 Object tracking

Zhang et al. [29] propose a tracking algorithm with an appearance model based on features extracted from multi-scale image feature space with data independent basis. Zhang et al. [28] model the motion of local patches of single object tracking that can be seamlessly applied to most part-based trackers in the literature. Zhang et al. [27] bound multiple Gaussian uncertainty to object tracking. Although these works show good algorithms for tracking the object, our framework focuses on the hand-held object. The hand-held object can be located by taking the advantage of the skeletal information from Kinect to track the hand position (it can be regarded as the object position).

2.3 Multi-view image recognition

Liu et al. [13] present multi-view Hessian discriminating sparse coding (mHDSC) which seamlessly integrates Hessian regularization with discriminating sparse coding for multi-view feature learning problems. Wu et al. [22] learn a multi-view low-rank dictionary for classification task. Both of the two works learn features on RGB images. Zha et al.

[26] propose to learn discriminating features from multiple views of RGB-D content, the feature learning function is formulated as a robust non-negative graph embedding function over multiple graphs in various views. Different from these works, we implement the multi-view feature learning on depth and RGB images using a unified deep learning framework, it can automatically learn different features from different modalities and dig out the complementary factors between the two modalities.

2.4 Hand-held object recognition

Hand-held object recognition can be divided into first-person interface and second-person interface. For the first interface works [17, 18, 23], [23] is driven by a head-pose calculation and laser pointer guidance to estimate the region of interest for the hand-held and object-at-distance. They compute the region of interest to recognize the hand-held object with SIFT. The difference in our work is that we use the location of hand to locate hand-held object which is more precise. [17] is a first-person view. They use motion to separate out hand-held object and combine the motion of object location and background movement as well as some temporal cues for a max-margin classifier. But this work is about first-person interfaces and our work is designed for second-person interface which has a wider application. Joes et al. [18] provide a dataset SHORT (The Small Hand-held Object Recognition Test), while the object covers most of the image, which makes the detection simpler, even without segmentation. This also makes the application restrictive.

Many works [12, 14, 15] focus on second-person interface. They focus on the scenario containing user, object and background. The main idea of these works is using RGB-D devices (i.e, Kinect, RealSense) to capture both the RGB and depth information, they take advantage of skeletal information of user to locate the hand position and depth information to segment the object. While the features they use are either all hand-crafted or partial hand-crafted, it's time-consuming and have a poor transformation to other datasets. Our work provides to uses CNN to learn features, which can sufficiently dig out the characteristics of the dataset and also adjust to other datasets.

3 The proposed framework

The framework consists of three parts: 1) object segmentation; 2) modality-specific feature learning; 3) high-level feature learning.

In the object region segmentation procedure, we use a method described in [15], which uses skeletal information and depth image to segment the hand-held object. The method first locates the hand position via skeletal information which can be acquired from Kinect. Then a region growing algorithm takes the average depth information computed from the hand position and its eight neighbor pixels as seed depth, if the difference between the neighbor pixel and the seed depth is in a predefined threshold, the neighbor pixel will be added to the region set, the procedure is repeated until no more neighbor pixel can be added. After obtaining the target region depth information, the RGB object region is directly segmented from the RGB image using the same bounding-box parameters with the depth object region in depth image.

After segmentation, we use the object regions including RGB and depth as input. The features on the object regions are obtained by the following procedures as illustrated in Fig. 2. Firstly, the feature on RGB image is learned on deep network (e.g. VGGNet [19]) and the depth feature is learned on shallow network (e.g. AlexNet [10]) respectively. Secondly,

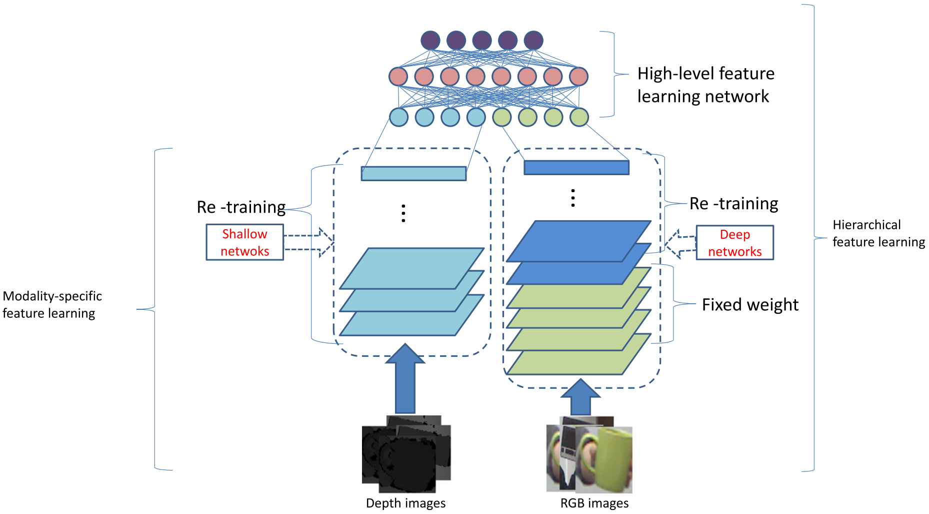

High-level feature learning network
Hierarchical feature learning
Re -training
Shallow netwoks
Re -training
Deep networks
Fixed weight
Modality-specific feature learning
Depth images
RGB images

Fig. 2 The proposed framework

we use high-level feature learning network to fuse the learned features on RGB and depth images and further learn a high-level and abstract feature. We use modality-specific and hierarchical networks (mshCNN) to denote our architecture.

4 Modality-specific feature learning

Because RGB-D hand-held object recognition contains different modal information (including RGB and depth) and the two modalities have a nontrivial difference such as appearance and color, we divide the feature learning procedure on RGB and depth images into model selection and learning strategies designing. We use modality-specific CNN (msCNN) to denote the feature learning procedure on RGB and depth images.

4.1 Models selection

Although RGB image and depth image are simultaneously captured by the RGB-D device, they describe different aspects of the object. The RGB image keeps rich color and texture information about the object region which can help to describe the line and object contour, while single RGB image can not describe the stereo information of the object. For example a yellow ball in RGB image will be presented as a yellow circle. On the contrary, the depth image has simple content (often represented by gray image) and loses rich color information of the object, while it keeps the depth information of each pixel on the object, which can describe the shape and spatial structure information about the object surface. Different modal information and different description aspects of the object make it hard to use one CNN model to learn the two modal features. At the same time, the depth information contains less information than the RGB image, the modal complexity for depth feature learning needs to be less than the RGB feature to avoid over-fitting.

For the above reasons, we implement two different models: a deep network for RGB feature learning and a shallow network for depth feature learning.

4.2 Learning strategies designing

Convolutional neural networks have been proven to have a good performance on the image classification [11, 19] and object detection [5]. While the CNN model often needs a large scale image dataset (such as ImageNet2012 with 1,000,000 images) for training. Current available hand-held object datasets such as HOD (it has about 12,800 RGB and depth image pairs in total) are too small that training a common and standard network model from scratch will lead to over-fitting. Matthew and Rob [24] show that different layers in CNN respond to different level features, the low layers learn the fundamental features, the high layers learn more abstract and class-correlated features. Taking AlexNet [10] as an example, the layer 2 responds to corners and other edge/color conjunctions, the high layers such as the layer 5 may learn the objects with significant pose variation. That is to say the parameters learned in low layer are sensitive to common and basic features, the parameters in high layer are specific to categories. The learned convolutional filters will be activated by basic features, which has less correlation with the dataset when the layer is low. Therefore, we use ImageNet2012 dataset to pretrain the networks to learn basic convolutional filters and initialize the fully connected layer parameters.

Because the pretrained networks are learned from the ImageNet2012 dataset, it can not perfectly describe the HOD dataset, especially for depth images. In order to generalize the initialized networks to the RGB-D dataset, we need to finetune the networks to target dataset. Because RGB images in HOD are more similar with the ImageNet2012 dataset in visual information (e.g. color and contour) than the depth information. As shown in Fig. 3, the images in first row are from HOD and the images in second row are from ImageNet2012. The first row and second row are examples of the same keyboard and cup categories. For RGB images in HOD, the keyboard images in HOD have very similar shape, text and color with the keyboard images in ImageNet2012. Although the cup images in HOD have different colors in the two datasets, they still have some common features in structure and shape.

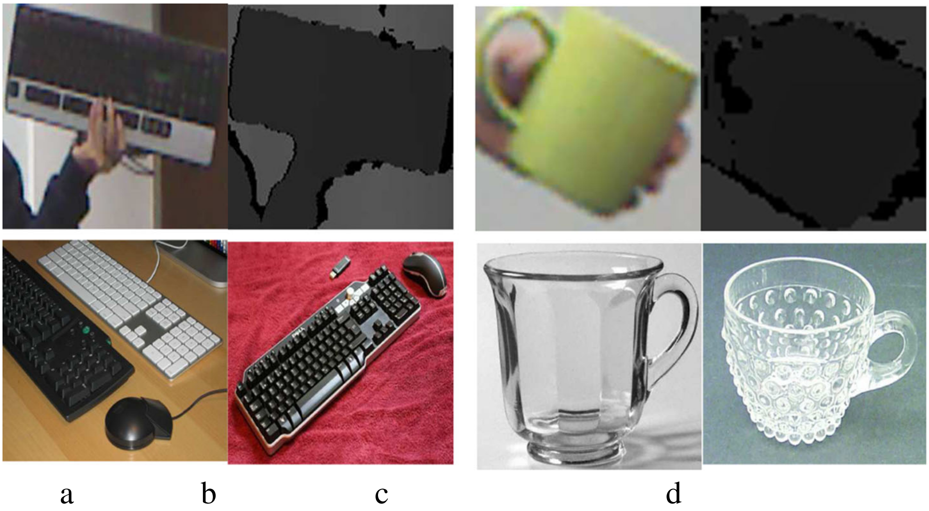


Fig. 3 Different dataset examples. The images in first row are from HOD dataset, the images in second row are from ImageNet2012 dataset

For example, they both have handle and rim. By contrast, the depth images in HOD have larger difference in appearance with the same categorical examples from ImageNet2012. For example, the keyboard example image, the depth image loses many detail information such as the key and layout on keyboard compared with the RGB image in ImageNet2012, but the depth image reserves the general contour information and the depth of each pixel. Therefore, we regard the depth as another description aspect of the object, which has a far distance in common appearance and pattern with the ImageNet2012 than the RGB images in HOD. Based on above analysis, we assume that: when the difference in two image set is large, it needs to have a full parameters finetuning on the original network; when the difference is small, we just need to finetune partial parameters on the original network (the low layers share common basic feature filters).

Therefore, different finetuning strategies are chosen for the two modal feature learning. For RGB image feature learning, we keep parameters in convolutional layers unchangeable to retain the learned basic feature filters, and only finetune the fully connected layers in a small learning ratio to make the fully connected layers adapt to the new dataset. For depth image feature learning, we finetune all parameters in the network to make low-layer convolutional filters can adjust to depth features (e.g. attaching more importance to general contour filters and less importance to color filters) and the fully connected layers parameters adjust to the finetuned low layers. The experimental results validate our strategies.

Except for implementing different finetuning strategies on different modalities, we consider that RGB and depth images have different content complexity. RGB images always contain more information than depth images, such as color and texture. As we all know that the content is more complex, there needs more parameters (such as increasing the filters in each layer and the layer number) to learn the pattern contained in the image, that may be one reason why there are many works that use deeper network [19, 20]. Considering different content complexities in RGB image and depth image, we use deep network for RGB feature learning and a relatively shallow network for depth feature learning.

5 High-level feature learning

Many works focus on directly concatenating multiple features together or assigning different weights to different features and then concatenating them together. While these methods ignore the relations among the features, for example the complementarity and noise. Directly combining two modal features is insufficient to have a high-level and comprehensive representation about the object region. The RGB feature and depth feature are learned separately in Section 4, they describe the same object and the features should be fused to learn a more comprehensive and high-level feature representation (like the functions of each layer in CNN). In order to learn a more discriminative feature for HOD, as shown in Fig. 2, we implement a high-level feature learning network to combine the previous learned features on RGB and depth. which can learn a more comprehensive and representative feature than single modal feature. The network is a fully connected network which takes the concatenated vector of RGB and depth features as input. Each hidden node in the first layer is connected with all the concatenated vector elements as input, this makes the output of hidden node can combine both depth and RGB features. With the nonlinear transformation, the network can learn a more representative feature for target dataset.

We use $R = [r_1, r_2, ..., r_n]$ to denote learned RGB feature, the dimension is n, $D = [d_1, d_2, ..., d_m]$ denotes the learned depth feature, the dimension is m.

Normalization depth feature and RGB feature are learned from different CNN models, which make them have different ranges in feature variable values. Simply inputting two modal features into the fully connected network will make the network itself have to adjust the weights on different amplitude in a unified learning ratio, this may lead to slowing down the weight learning procedure and influencing the final performance as well. Therefore, we implement normalization on both depth and RGB features to restrict them between -1 and 1. The normalization is shown in (1). The x_{max} and x_{min} are the maximum and minimum of all features in one dimension. y_{max} and y_{min} are the maximum and minimum of target variable range, x_i is one dimensional value of i-th feature, x_i' is the corresponding value after normalization.

$$x_i' = (y_{max} - y_{min}) . \frac{(x_i - x_{min})}{(x_{max} - x_{min})} + y_{min} \tag{1}$$

After normalizing the feature, the two features are fed into the network. For each node in the first fully connected layer, its output can be formulated as follow:

$$x_z = f\left(\sum_{i=1}^{n} w_i r_i + \sum_{j=i+1}^{m} w_j d_{j-i}\right) \tag{2}$$

where $f(\cdot)$ is an activation function, w_i is the weight of i-th input node, r_i is i-th dimensional value in RGB feature, d_j is j-th dimensional value in depth feature, x_z is output of z-th hidden node. For each node in the second layer, the input contains all the outputs from previous layer. In this way, each node of the next layer contains both RGB and depth feature information, meanwhile it can learn the weights on each dimension of both RGB and depth features. Since the first layer is similar to feature selection and re-weighing the input feature, therefore we implement two layers fully connected network. Twice nonlinear transformation and jointly learning weights on RGB and depth features make the final feature have an abstract and comprehensive representation of the object region. Different weights can find the complementarity and suppress the noise in RGB and depth features.

Many works [12, 14, 15] use SVM to classify the fused feature. The reason is that they have hand-crafted features which make them can not combine the feature fusion procedure and classification together, meanwhile SVM performs better than softmax in their works. In this work, we automatically learn a softmax classifier (it can be used to compute the loss in training procedure and classify in test procedure) when training the networks, this avoids spending too much time to train SVM. The experimental results validate that in our unified model the softmax also has a good performance compared with SVM.

6 Experiments

6.1 Dataset and experimental setup

Dataset we use HOD [14] dataset for evaluation. HOD dataset is the only dataset for RGB-D hand-held object recognition task, which consists of 16 daily categories and 4 instances for each category. For each category, there are 800 RGB and depth image pairs and 200 pairs for each instance.

Evaluation the evaluation consists of two tasks: *seen* and *unseen*. *Seen*: training and test sets both contain images from all the instances. *Unseen*: the training set contains images from instance 1, 2 and 3, the instance 4 is used as test set.

6.2 Evaluations on different feature learning models

For convolutional neural network, different layers have different representation level [25], when the layer is higher, the learned feature is more specific to the object. That means low layer feature has a high probability in sharing low-level feature filters across different image datasets. In order to have a good insight into the influence of finetuning on RGB and depth images, we use *unseen* data to try different adjustment strategies as shown in Fig. 4. An eight layers AlexNet model is used, which consists of five convolutional layers and three fully connected layers. The model is pretrained on ImagetNet2012 dataset, which contains 1000 categories and about 1,000,000 images in total. Because the convolutional layers are basic feature filters, we take them as a whole for finetuning instead of finetuning each layer.

In Fig. 4, the whole network finetuning on RGB images has the worst performance compared with the other three finetuning strategies, it is even worse than directly extracting feature using pretrained model without finetuning. This indicates that using the RGB images of HOD dataset to finetune the convolutional filters may impact or even damage the representation and discrimination ability of the filters. The filters trained on an enormous and diverse RGB image dataset (ImageNet2012) already have good descriptions about basic RGB image features like color and contour, and the RGB images of HOD are few and the image content are very simple as shown in Fig. 3. While finetuning is a parameter rectification procedure, simple and biased training data will make the convolutional filters focus on part of the filters and reduce the weights of other filters in the finetuning, which weakens the final network in a comprehensive feature description of the object. On the contrary, finetuning on the last fully connected layers from 6 to 8 layers, the three finetuning strategies all have significant improvement compared with the whole network finetuning. They also outperform the performance that using the pretrained model to extract feature without finetuning. This is because convolutional layers have a good description about low-level feature, while the fully connected network is more sensitive to new dataset. It's interesting that

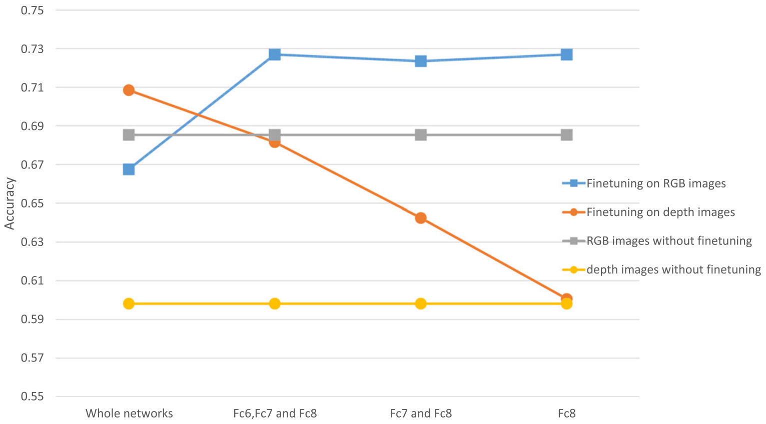


Fig. 4 Different finetuning strategies. We compare the finetuning performance on different modal images(RGB and depth). The horizontal axis shows the different fintuning strategies. "Whole networks" means we fintune the whole AlextNet. "Fc6, Fc7 and Fc8" denotes finetuning three fully connected networks, "Fc7 and Fc8" denotes finetuning the latter two fully connected networks and "Fc8" denotes only finetuning the last fully connected network. "without finetuning" means that we use a pretrained model to extract features

the performance on finetuned 6-8 fully connected network is almost the same with the finetuned 8-th layers, and they both perform better than finetuned 7-8 layers, this might because the 8-th finetuning has a high-level feature input from 7-th layer. While the 6-8 layers can learn a more specific feature to HOD dataset, the 6-th layer feature are not as high-level as the 7-th layer, this makes the 7-8 layers finetuning can not learn a good representation of the HOD dataset.

For depth images, the whole network finetuning outperforms the other finetuning methods, and the performance declines with reducing the finetuning layers. This illustrates that the model needs to finetune the parameters from convolutional layers to fully connected layers to make the model be able to adjust to depth images. The main reason is that depth images have great difference in appearance with RGB images as shown in Fig. 3. The convolutional filters pretrained on ImageNet2012 can not transfer all the basic feature filters to depth images (such as the filters which will be activated by color information will be useless for depth image, the filters which will be activated by texture information will also not work), this makes the convolutional filters can not extract a representative feature from object depth image. Therefore, the fully connected layers which take the convolutional results as input will also be influenced. From another perspective, the whole networks finetuning makes the pretrained model adjust to the depth data and improve the representation ability to the new depth images. The feature directly extracted from the pretrained model has the worst performance, it's almost similar to the result of only finetuning the 8-th layer, this is because the pretrained model has a great difference with depth images in the HOD dataset.

The evaluations on different architectures assist us finding suitable and discriminating modality-specific feature learning methods for both RGB and depth images from HOD dataset. For depth images we need to finetune the whole networks. For RGB images, the network has already learned common filters, we need to finetune the fully connected layers to avoid new data bringing noise to the filters.

6.3 Evaluations on different high-level feature learning methods

As shown in Fig. 4, we implement whole network finetuning on depth images and fully connected layers finetuning on RGB images. For the evaluation of our high-level feature learning method, we select different networks and low-level feature sources.

Besides the different fusion networks, the AlexNet is replaced with VGGNet [19] on RGB feature learning, this is mainly because that the VGGNet which has 16 layers has been proved to have a better representation ability than AlexNet, the fully connected layers of the two models have same architectures. The number of convolutional layers of VGGNet is more than AlexNet, which makes the networks have better representative ability than AlexNet. Therefore, we choose VGGNet for RGB images feature learning. For depth images, because the content in depth image is simple, the VGGNet may benefit RGB images, but its filters can not easily adjust to the depth images. Thus, we implement the AlextNet for depth images feature learning and finetune the whole network.

We compare different feature fusion methods under different combination. The fusion methods include using two layers fully connected networks and three layers fully connected networks. The combination methods include: 1) using AlextNet on both RGB and depth images (using "8C8D" to denote); 2) using AlexNet on depth images and VGGNet on RGB images (using "16C8D" to denote). We use "FT" to denote the model which uses finetuning and "without FT" means not using finetuning which directly uses the pretrained model to extract features from RGB and depth images. As shown in Fig. 5, the performance of

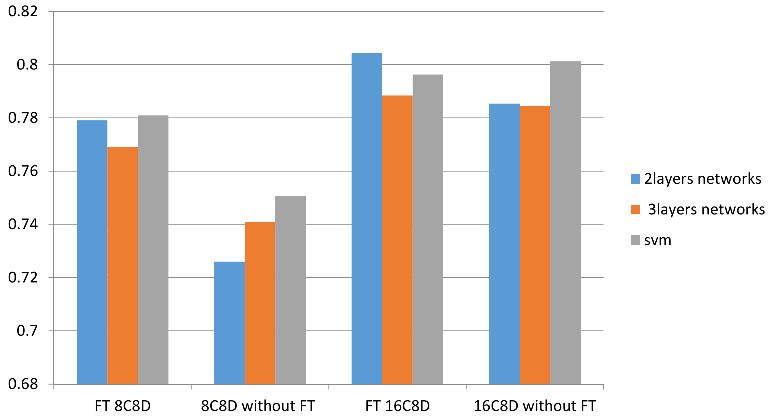


Fig. 5 Different fusion and combination methods. The horizontal axis denotes different combination strategies. "8C8D" means 8 layers networks on RGB and depth images. "FT" means using finetuning. "16C8D" means we use16 layers networks (VGGNet) on RGB images and 8 layers networks (AlexNet) on depth images. For different fusion strategies, "2 layers networks" means two layer fully connected networks, "3 layer networks" means three layer fully connected networks, "svm" means that we directly concatenate the two modal features and use svm for classification

"16C8D" is better than "8C8D" on both finetuning and no-finetuning conditions, this validates that the VGGNet has a better description for RGB images. Besides, comparing "FT 8C8D" and "8C8D without FT" and comparing "FT 16C8D" and "16C8D without FT" show that the combination of different modal features under finetuning performs better than using pretrained models, this illustrates that finetuning can help to improve the performance on both "8C8D" and "16C8D". For fusion methods, high-level feature learning networks using two layers networks perform better on both finetuned "8C8D" and "16C8D" than the three layers, even for the no-finetuned "16C8D". This is probably because that the HOD dataset is small and three layers have too many parameters, which makes the network model too complex for the data being over-fitting. For the "8C8D without FT", it's not finetuned on the HOD and have less parameters, so this makes the network not over-fitting.

We also compare our method with SVM results, which are obtained by directly replacing the fully connected layers with SVM. The results of SVM nearly outperforms all the performance using fully connected layers. This is because that SVM has been proved to be an outstanding classifier in many fields, which has a good performance. The best performance is achieved by using finetuned "16C8D" with two layers networks, which is the only one that outperforms the result using SVM, it shows that finetuning on "16C8D" and two layers fully connected networks can have a significant improvement in the feature learning and representation.

6.4 Evaluation on the proposed approach

In this experiment, we compare the proposed framework with previous works on hand-held object recognition. Table 1 illustrates the different methods from different works, the accuracy is obtained on $Seen$ and $Unseen$ conditions. It is obvious that the proposed method gets best performance on $Unseen$ condition compared with the rest methods. This means that the proposed modality-specific and hierarchical feature learning method can improve the visual description ability of RGB-D objects in HOD dataset under the $Unseen$ condition. While

Table 1 Experimental results.RGB_{16l}+D_{8l} denotes using pretrained 16-layer VGGNet and 8-layer AlexNext to extract the features on RGB and depth images separately

Method	Classifier	Accuracy (%)	
		Seen	Unseen
$C - CNN + D - CNN + C3 + ESF + GRSD$ [14]	SVM	83.33	69.50
$C3 + ESF + GRSD$ [14]	SVM	70.05	58.56
$MKL(C - CNN + D - CNN + C3 + ESF, C - CNN + D - CNN, C3, ESF)$ [14]	SVM	85.10	75.31
$MKL(C - CNN, D - CNN, C3, ESF, GRSD)$ [14]	SVM	**88.59**	73.31
RGB_{16l}+D_{8l}	SVM	82.09	80.13
$msCNN$	SVM	84.86	79.63
$mshCNN$ (our method)	softmax	84.24	**80.44**

for the *Seen* condition, our approach does not have such improvements. The approach uses fully connected layers to learn high-level feature. This may be because that the *Seen* test images have the same instances appeared in the training data, which makes the features from test images are similar to the features from same instances in the training data. The nonlinear transformation in fully connected layers will change the features significantly, which may damage the similarity. Our approach also outperforms the RGB_{16l}+D_{8l} (it denotes directly extracting features on RGB and depth images using VGGNet and AlextNet.) method, this validates the effectiveness of our learning strategy. Our approach behaves better in *Unseen* condition than $msCNN$, this illustrates that our method is robust to *Unseen* data. Although $MKL(C-CNN, D-CNN, C3, ESF, GRSD)$ [14] obtains a very high performance in *Unseen* data, it uses five features for object region representation and our approach gets about 7 % improvement in *Unseen* data.

Besides the performance, softmax classifier is often simultaneously trained with the networks, this makes the training and test procedure more convenient and have a lower time and space complexity than using SVM. SVM classifier divides the procedure into two parts both in training and testing: first, we need to train the networks and extract the features; then a SVM is trained and used for prediction. In the computing efficiency, Lv et al. [15] extract ESF, $GRSD$ and $C3$ separately using CPU and extract $C-CNN$ as well as $D-CNN$ using GPU, which makes the feature extraction time-consuming. Besides, they concatenate all the features and use a SVM for classification, this needs additional time. Different from the above work, our method incorporates feature extraction and label prediction into a unified framework. Besides our method is under CNN architecture and implemented on GPU for acceleration. Therefore, the method is more efficient than the methods in [15]. In the experiment, we use K40 GPU for training and test, the test time of an image is about 0.2 s, which is more efficient than the method in [15] (the time of $MKL(C-CNN+D-CNN+C3+ESF, C-CNN+D-CNN, C3, ESF)$ is about 1.0s).

7 Conclusion

In this paper, we propose a novel hierarchical feature learning method under the setting of CNN for RGB-D hand-held object recognition. In the first step, the method implements networks with different depths for RGB and depth images, and uses modality-specific learning strategies to learn features, which can sufficiently dig out the characteristics of RGB and depth patterns as well as learn an adaptive and representative feature on each modality. In the second step, a high-level feature learning network is used to learn a more comprehensive feature from the RGB feature and depth feature. It can fuse different modal features as well as learn a high-level and categorical feature representation. The experimental results validate the efficiency of our method. In the future, we will extend this work in more generic tasks (e.g. RGB-D scene classification and RGB-D object classification).

Acknowledgments This work was supported in part by the National Basic Research 973 Program of China under Grant No. 2012CB316400, the National Natural Science Foundation of China under Grant Nos. 61532018, and 61322212 the National High Technology Research and Development 863 Program of China under Grant No. 2014AA015202. This work is also funded by Lenovo Outstanding Young Scientists Program (LOYS).

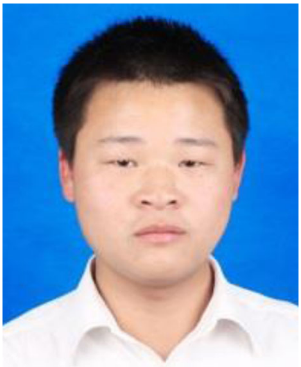
References

1. Beck C, Broun A, Mirmehdi M, Pipe A, Melhuish C (2014) Text line aggregation. Int Conf Pattern Recogn Appl Methods (ICPRAM), pp 393–401
2. Bo L, Ren X (2011) Depth kernel descriptors for object recognition. In: IROS, pp 821–826
3. Chai X, Li G, Lin Y, Xu Z, Tang Y, Chen X, Zhou M (2013) Sign language recognition and translation with kinect. In: ICAFGR
4. Fu Y, Cao L, Guo G, Huang TS (2008) Multiple feature fusion by subspace learning. In: Proceedings of the 2008 international conference on Content-based image and video retrieval. ACM, pp 127–134
5. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, pp 580–587
6. Gupta S, Arbeláez P, Girshick R, images JM (2014) Indoor scene understanding with rgb-d Bottom-up segmentation, object detection and semantic segmentation. IJCV, pp 1–17
7. Gupta S, Girshick RB, Arbelaez P, Malik J (2014) Learning rich features from RGB-d images for object detection and segmentation. CoRR, abs/1407:5736
8. Ji S, Xu W, Yang M, Yu K (2013) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
9. Kanezaki A, Marton Z-C, Pangercic D, Harada T, Kuniyoshi Y, Beetz M (2011) Voxelized shape and color histograms for rgb-d. In: IROS Workshop on Active Semantic Perception. Citeseer
10. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: NIPS, pp 1106–1114
11. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp 1097–1105
12. Liu S, Wang S, Wu L, Jiang S (2014) Multiple feature fusion based hand-held object recognition with rgb-d data. In: Proceedings of International Conference on Internet Multimedia Computing and Service. ACM, pp 303–306
13. Liu W, Tao D, Cheng J, Tang Y (2014) Multiview hessian discriminative sparse coding for image annotation. Comput Vis Image Understand 118:50–60
14. Lv X, Jiang S-Q, Herranz L, Wang S (2015) Rgb-d hand-held object recognition based on heterogeneous feature fusion. J Comput Sci Technol 30(2):340–352
15. Lv X, Wang S, Li X, Jiang S (2014) Combining heterogenous features for 3d hand-held object recognition. In: Proceedings SPIE, Optoelectronic Imaging and Multimedia Technology III, vol 9273, pp 92732I–92732I–10
16. Marton Z-C, Pangercic D, Rusu RaduB, Holzbach A, Beetz M (2010) Hierarchical object geometric categorization and appearance classification for mobile manipulation. In: Proceedings of the IEEE-RAS International Conference on Humanoid Robots, TN, USA
17. Ren X, Gu C (2010) Figure-ground segmentation improves handled object recognition in egocentric video. In: CVPR, pp 3137–3144
18. Rivera-Rubio J, Idrees S, Alexiou I, Hadjilucas L, Bharath AA (2014) Small hand-held object recognition test (short). In: WACV, pp 524–531
19. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409:1556
20. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2014) Going deeper with convolutions. CoRR, abs/1409:4842
21. Wohlkinger W, Vincze M (2011) Ensemble of shape functions for 3d object classification. In: ROBIO, pp 2987–2992
22. Wu F, Jing X-Y, You X, Yue D, Hu R, Yang J-Y (2016) Multi-view low-rank dictionary learning for image classification. Pattern Recogn 50:143–154
23. Xu RYD, Jin JS (2006) Individual object interaction for camera control and multimedia synchronization. In: ICASSP, vol 5
24. Zeiler MD, Fergus R (2013) Visualizing and understanding convolutional networks. CoRR, abs/1311:2901
25. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: ECCV, pp 818–833
26. Zha Z-J, Yang Y, Tang J, Wang M, Chua T-S (2015) Robust multiview feature learning for rgb-d image understanding. ACM Trans Intell Syst Technol, vol 6, pp 15:115:19

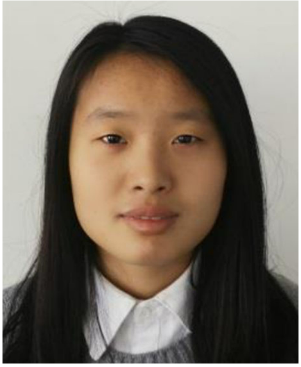
27. Zhang B, Perina A, Li Z, Murino V (2016) Bounding multiple gaussians uncertainty with application to object tracking. IJCV
28. Zhang B, Li Z, Perina A, Del Bue A, Murino V (2015) Adaptive local movement modelling for object tracking. In: WACV, pp 25–32
29. Zhang K, Zhang L, Yang M-H (2014) Fast compressive tracking. IEEE Trans Pattern Anal Mach Intell 36(10):2002–2015

Xiong Lv received his Bachelor's degree in Computer Science and Engineering degree from Beihang University, China in 2013. He is currently a graduate student of the Institute of Computing Technology of the Chinese Academy of Sciences, Beijing, China in 2015. His research interests include image understanding, human-system interaction with image, 2D and 3D object recognition.

Xinda Liu received the B.S. degree in school of Geometrics Engineering, China University of Mining & Technology, Beijing, China, in 2013. He is a Master degree candidate in Computer software and theory in the Ningxia University, Ningxia, China. His research interests include image processing, data mining, computer vision, and machine learning.



Xiangyang Li received the M.S. degree from the College of Information and Engineering, Capital Normal University, Beijing, China, in 2015. He is a Ph.D. student in computer science at the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include large-scale image classification, joint learning with language and vision, computer vision, and pattern recognition.



Xue Li received the B.S. degree in school of ShanDong University of Science and Technology, ShanDong, China, in 2014. She is a joint training postgraduate student in computer science at ShanDong University of Science and Technology and the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, China. Her research interests include image processing, incremental learning, and computer vision.

Shuqiang Jiang (SM'08) is a professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 100 papers on the related research topics. Dr. Jiang was supported by the New-Star program of Science and Technology of Beijing Metropolis in 2008. He won the Lu Jiaxi Young Talent Award from Chinese Academy of Sciences in 2012, and the CCF Award of Science and Technology in 2012. He is the senior member of IEEE, member of ACM, CCF, and YOCSEF. Prof. Jiang is the executive committee member of ACM SIGMM China chapter. He has been serving as the guest editor of the special issues for PR and MTA. He is the program chair of ICIMCS2010, special session chair of PCM2008, ICIMCS2012, area chair of PCIVT2011, publicity chair of PCM2011 and proceedings chair of MMSP2011. He has also served as a TPC member for more than 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICIP, and PCM.

Zhiqiang He joined Lenovo Group in 1986 and is currently the Senior Vice President of the Company and President of the Ecosystem and Cloud Services Business Group. This group is responsible for building Lenovos ecosystem and customer relationship through cloud services, as well as exploring and driving growth in the broader personal and enterprise Internet services. Previously, Mr. He was the Chief Technology Officer and held various leadership positions in Lenovo, particularly in overseeing Lenovos Research & Technology initiatives and systems. Mr. He is doctoral supervisor at the Institute of Computing Technology of Chinese Academy of Sciences and Beihang University. Mr. He holds a bachelors degree in computer communication from Beijing University of Posts and Telecommunications and a masters degree in computer engineering from the Institute of Computing Technology of Chinese Academy of Sciences.