RGB-D Object Recognition from Hand-Held Object Teaching

Leixian Qiao
Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences Institute of Computing Technology Chinese Academy of Sciences
Beijing,China
leixian.qiao@vipl.ict.ac.cn

Xue Li
Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences Institute of Computing Technology Chinese Academy of Sciences
Beijing,China
xue.li@vipl.ict.ac.cn

Shuqiang Jiang
Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences Institute of Computing Technology Chinese Academy of Sciences
Beijing,China
sqjiang@ict.ac.cn

ABSTRACT

For RGB-D object recognition, conventional methods only focus on classification, which neglects the importance of humans for object segmentation and object concept learning in the interaction and has limitations when transferring the learned knowledge to general indoor scenes. In this paper, we propose a system that humans can teach robots object concepts and instances by the way of hand-held object recognition, a newly research topic in RGB-D object recognition, then these concepts and instances are transferred to more general indoor scenes to recognize target objects. Unlike traditional approaches of RGB-D object recognition, human-machine interaction is considered in our system, in which robots can obtain the object region from object segmentation and enhance its own interactive learning ability. In addition, we propose an RGB-D object dataset to match the hand-held object dataset. In experiments, we use CNN to learn feature representation of RGB-D images in the hand-held training object dataset and RGB-D test object dataset respectively. Experimental results show that our system has a strong capability in RGB-D object recognition.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

General Terms

Design, Experimentation

Keywords

Hand-held object recognition, RGB-D object recognition, Convolutional neural network

1. INTRODUCTION

Although RGB-D object recognition develops rapidly in computer vision and robotics with the popularization of high-quality RGB-D cameras, like Kinect, ASUS Xtion Pro Live, recognition process in previous work lacks people involvement. Traditional methods merely focus on how to learn better feature representation and get a more accurate classification result. However, they neglect the influence that humans contribute to acquiring object regions from RGB-D cameras and learning the knowledge concerning object concepts from speech in early image processing, in addition, it is uncertain that the system can be robust when transferring the learned knowledge of object concepts and instances to other indoor environments. To address this problem, in this paper, we propose an RGB-D recognition system which can learn object concepts from hand-held object recognition in CNN framework, then transfer the learned knowledge to new indoor scenes to search and recognize the same or similar objects.In hand-held object recognition, the objects in hands can be acquired from object segmentation and these object concepts and instances can be also mastered. Our system combines learning from hand-held teaching and transferring the learned knowledge to real-world surroundings successfully, which especially accords with robot learning activities and embodies a complete flow of robots mastering learned object concepts and instances.

For object recognition, its goal is to learn robust feature representation and train optimal classifiers to gain strong classification results. To learn feature representation, conventional methods use hand-crafted features, such as SIFT[12], SURF[2]and HOG[17] to describe images from distinct sides such as color, texture, edge, shape, etc. However, hand-crafted features have some restrictions for image characterization, which only adapt to specific small dataset and cannot characterize high-level semantics representation. With the development of deep learning, deep neural networks are applied to image representation gradually, such as CNN[10], Autoencoder[11]and RNN[13]. Features can be learned by deep neural networks without supervision and the learned features are robust and high-level for image classification. CNN has a strong ability in image characterization and has been applied to RGB-D object recognition. In this paper, we utilize CNN framework to extract RGB-D image features.

Hand-held object recognition[14, 15] is an important part in interactive learning in vision-based environment. It reflects a process of teaching and learning. It can be imagined that the robot can continue improving its own learning ability when seeing objects held in the user's hands and hearing user's speech describing the objects. With the popularization of RGB-D camera, not only is color information obtained, but also human body information and depth data

ICIMCS '16, August 19-21, 2016, Xian, China
© 2016 ACM. ISBN 978-1-4503-4850-8/16/08…$15.00
DOI: http://dx.doi.org/10.1145/3007669.3007713

around the surrounding scene can be easily provided. Thus, robots can generate a better understanding to users and environments. In hand-held object recognition, hand positions can be regarded as initial points due to objects in the user's hands, and they can be acquired from the user's skeletal information, so it is easy to separate the objects in hands according to depth cues.

In this paper, we provide an RGB-D object dataset which has same categories and instances as the hand-held object dataset(HOD) [14]. In experiments, we obtain segmentation regions of target objects from RGB images and depth maps in HOD and RGB-D object dataset. Next, we use HOD as the train set to learn object concepts and instances for the task of hand-held teaching, then RGB-D object dataset is used as the test set, which evaluates the recognition capability of our system in general indoor scenes. It can be indicated from the experimental results that our system attains the goal of learning from hand-held teaching and recognizing objects in new indoor surroundings.

The rest layout of this paper is as follow: Section II is some related work on RGB-D recognition, Section III describes the system framework, Section IV shows experiment setup and experimental results. Finally, Section V concludes the paper.

2. RELATED WORK

In recent years, RGB-D object recognition has been a booming research field in computer vision and robotics with the widely utilization of cheap RGB-D sensors. Some typical methods extract hand-crafted features for RGB-D object recognition. [5] proposed a HMP method based hierarchical sparse coding to learn local features of RGB images and depth maps. [4] also presented a depth kernel descriptor to fuse multiple hand-crafted features for the mission of object recognition. [3] provided a convolutional k-means descriptor for RGB and depth images, and it can learn meaningful local feature description automatically. Besides, With the rise of deep learning, Deep neural networks gain roaring success in computer vision tasks, and most of methods of RGB-D object recognition extract features in the framework of deep neural networks, taking the advantage of well performing feature learning. [16] presented a method combining convolutional and recursive neural networks to learn features and classify the RGB-D images, in their framework, RGB and depth features are learned by CNN-RNN individually. The final feature are represented by the concatenated feature vector. [7]proposed an approach which encodes the depth image with three channels at each pixel: horizontal disparity, height above ground, and the angle between the local surface normal and the inferred gravity direction, then train CNN on respective channels. [8] proposed a deep Regularized Reconstruction Independent Component Analysis network (R^2ICA) for RGB-D image classification, which builds the relationship between the gray-scale and depth images corresponding to the same object or scene and use local contrast normalization and spatial pooling to learn robust image representation. [1] divides instances in datasets into different subsets according the object shape, then integrates sparse autoencoder and recursive neural network to learn a fused feature by concatenating the RGB feature and the depth feature.

Hand-held object recognition is a newly research area in RGB-D object recognition, [14] proposed an approach which assembles RGB and depth features learned by pretrained CNN and 3D point cloud features into a total feature vector and gains a good recognition result in seen and unseen category recognition. [15] proposed a hierarchical feature learning framework, which has different network architectures and learning manners in the RGB modality and the depth modality, then the features are input to the high-level neural network to learn robust feature representation, which gets an semantic description in the final fused feature vector.

3. THE PROPOSED SYSTEM

In our system, we want to realize the goal that humans can teach robots object concepts and instances by hand-held teaching, then robots can search and recognize target objects in other indoor scene using the prior learned knowledge . The pipeline of our system is shown in Figure 1. Our system can be divided into three parts: object segmentation, feature extraction and image classification. In the part of segmentation, we use region-growing algorithm to process hand-held RGB-D images based on depth maps and human skeletal information. For images in the RGB-D object dataset, we use the approach of interactive region growing to gain the target object mask. After segmentation, we utilize CNN framework to learn feature representation. In Figure 1, we can find that our network consists of two streams processing RGB images and depth maps independently and the features in the FC7 layers of two modalities are concatenated as the discriminative feature. After that, rbf kernel SVM will be trained to classify the final combined feature. In our system, to represent the principle of teaching and learning, we firstly learn knowledge concerning object concepts and instances in pretrained CNN on segmented images in HOD. When obtaining the knowledge, it is proved that robots have learned these object concepts and instances. Then, we use the segmented images in the RGB-D object dataset as the test set to evaluate the classification ability. Then we will explain the segmentation and feature learning in detail.

3.1 Object Segmentation

For the HOD dataset, we separate objects from the depth map. First of all, the depth map is filtered to eliminate image noise and a portion of black holes are filled. When the depth map is captured, body skeletal data is also collected by RGB-D sensors which assists in locating positions of target objects. Hand-held object recognition meets the fact that target objects are usually in the left or right hand, so the initial segmentation points are the position of double hands in depth maps. Next, using double hand position as initial seeds, we can obtain the object mask in region-growing algorithm, which clusters the similar set of points less than the depth threshold. The algorithm makes good use of spatial information in the depth map, and it is simple and practical for object segmentation in the hand-held object scene. The detailed description of region growing algorithm are shown in [14].

For RGB-D object dataset, we use the interactive region-growing method to obtain the target object mask. The initial point is manually chosen from the target object region in the depth map, and used as the initial seed, then the object mask is acquired by clustering the points belongs to the target object. Meanwhile, the segmented region in the RGB image and the depth map are also obtained.In the future, we will use selective search algorithm to get the target object proposal from the RGB image based on different slices of depth value in the depth map.

3.2 Feature Learning

In this paper, the deep feature is used to represent the RGB image and the depth map. We utilize CaffeNet[9, 10] to extract features from RGB images and depth maps respectively. Before extracting features, we need to preprocess RGB images and depth maps for network requirements. For both RGB and depth modalities, the input images are resized to 227×227×3 pixels and each image removes the mean value of the image. For extracting features, firstly, we represent image features on feed-forward CaffeNet, which is pretrained on a large dataset(ISLVRC 2012 for ImageNet) and re-

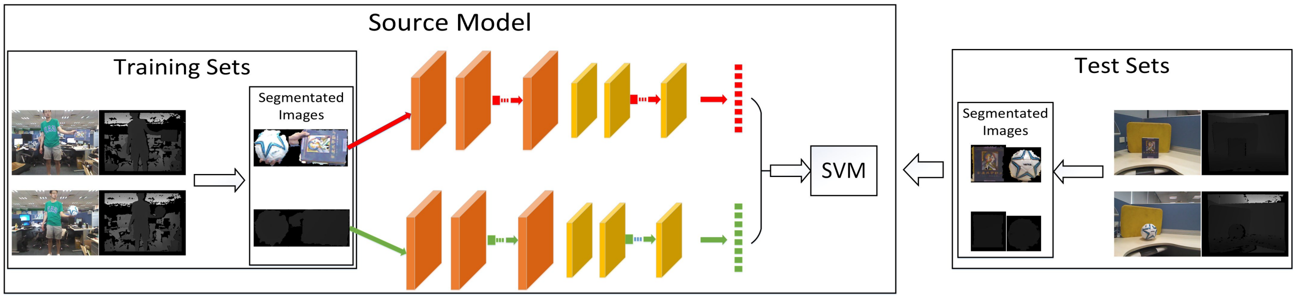

Figure 1: The pipeline of our system.The system includes 3 parts: object segmentation, feature extraction and image classification

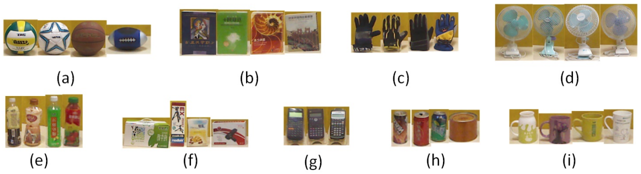

Figure 2: The RGB-D object dataset proposed in the paper, which includes 9 categories:(a)ball, (b)book, (c)glove, (d)fan, (e)bottle, (f)box, (g)calculator, (h)can and (i)cup. In addition, we extend the instances in 9 same categories in HOD to keep consistent with the RGB-D object dataset.

trained on HOD and proposed RGB-D object dataset. Then we extract the 4096-dimensional feature of the FC7 layer from the RGB stream and the depth stream independently. In addition, the features learned from both of the RGB modality and the depth modality are concatenated into an 8192-dimensional feature vector and used as final image representation.

4. EXPERIMENTS

4.1 Datasets

4.1.1 Hand-Held RGB-D Object Dataset

For hand-held teaching task, we use the hand-held RGB-D object dataset(HOD) in [14]. The dataset extracts 12800 video frames from Kinect camera placed 1.5m above the ground. In each frame, we obtain an RGB image, a depth map and human skeletal data from color and depth stream in Kinect. For reducing image noise and gathering stable skeletal data during the capturing process, human body in front of Kinect should keep a suitable distance, approximately 1~2m. The hand-held object dataset contains 16 common object classes, each of which has 4 instances, 64 instances totally. Every instance are captured by two person in two different scenes. In our experiments, we choose 9 categories from HOD, and extend some instances to meet the requirement of the RGB-D object dataset.

4.1.2 RGB-D Object Dataset

For the task of classification in general indoor scenes, we collect an RGB-D object dataset captured by Kinect in the indoor scene.The dataset includes 9 same everyday object classes as HOD, which are ball, book, bottle, box, calculator, can, cup, fan and glove. Each of category contains 3 or 4 instances. The total number of the RGB-D object dataset is 6602. The dataset gains the RGB images and depth maps from multiple view angles around the horizon and vertical axis.The dataset is shown in Figure 2.

Note that the same category among two datasets has same instances, and the distributions of two datasets are consistent with each other.

4.2 Experiment Setup

Our system is employed on CaffeNet framework which consists of five convolutional layers and three fully-connected layers. For RGB and depth modalities, both of them utilize the same framework pretrained on ImageNet 2012. We preprocess images according to Section 3.2 for both of the RGB and depth modality, then the processed RGB images and depth maps are independently input to the network and the FC7 feature for CaffeNet is extracted as the image feature. Note that the FC6 and FC7 of two streams are followed by a dropout layer with a dropout rate of 0.5. Finally, we concatenate the features of two modalities into a final feature vector, and then we use the rbf kernel to train a classifier in LibSVM[6].

Towards two datasets, we evaluate the performance of our system on two levels: seen object recognition and unseen object recognition. For seen object recognition, just as its name implies, it represents that both of the train set and the test set have same instances in every category. Even though the images in each instance among train and test sets are different from poses, views and light changes, in the aspect of feature representation, they still keep some similarities with each other, so it is easy for the system to recognize the object in the test image. In the experiment, we train all of instances in HOD and test all of instances in the RGB-D object dataset. For unseen object recognition, although some object categories have been seen in advance, it is difficult for system to recognize the images from some new instances in the est set which don't appear in the train set. In the experiment, we choose one instance in every category to form the test set, and the remaining instances in every category are added into the train set.

4.3 Experimental results

We evaluate seen and unseen experiments in RGB, depth and concatenated features respectively. Table 1 shows experimental results of seen and unseen object recognition in three different kinds of features,and it reveals that the classification accuracy of using RGB+Depth fused features is higher than 2~4% than using RGB feature, and it can represent images better than individual RGB or depth features regardless of unseen and seen object recognition.

Besides, the performance on seen object recognition is very robust even if the segmented images have interference of inevitable hand information from Table 1 and Figure 3. For unseen object recognition, when using concatenated features, the classification accuracy reaches up to 61.88% and it illustrates that our system has good abilities of recognizing unseen instances in prior trained knowledge of other instances from Table 1 and Figure 3 although our dataset are not big enough.

Table 1: Classification accuracy in RGB, Depth and RGB+Depth fused features

Accuracy(%)	RGB	Depth	RGB+Depth
Seen recognition	82.61	51.70	86.35
Unseen recognition	59.58	40.27	61.88

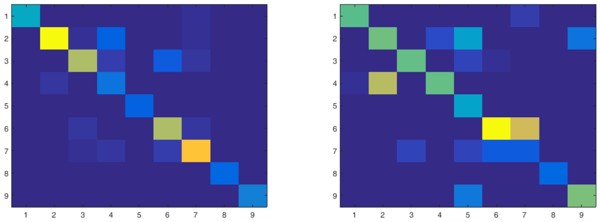

Figure 3: The respective confusion matrices of seen and unseen object recognition in the concatenated feature

In a word, our system achieves the goal of learning from hand-held teaching and transferring the learned knowledge to the new indoor scene, and performs well in seen and unseen object recognition.

5. CONCLUSION

In this paper, we propose an system which integrates modules of learning from hand-held teaching and application in the general indoor scene. The system conforms to the learning behaviors of robots, the modality of hand-held object recognition achieves the goal of teaching robots object concepts and instances, then all of them are mastered by robots. In practice, robots utilize mastered knowledge to recognize objects in more general indoor scenes. In our system, we separate RGB images and depth maps in HOD and RGB-D object dataset proposed in the paper, and CaffeNet is pre-trained on ImageNet 2012 to extract features on the RGB and depth stream, finally, the RGB and the depth feature are concatenated to represent the object. In experiments, we learn object concepts and instances from HOD, and the RGB-D object dataset is used to test the classification capability in new indoor environments. Experimental results show that the performance on seen and unseen object recognition is wonderful and the proposed system is capable of learning from hand-held teaching and transferring learned knowledge to the new indoor scene.

Acknowledgments

This work was supported in part by the National Basic Research 973 Program of China under Grant No. 2012CB316400, the National High Technology Research and Development 863 Program of China under Grant No. 2014AA015202, the National Natural Science Foundation of China under Grant No. 61532018 and 61322212, Beijing Science And Technology Project under Grant No. D161100001816001. This work is also funded by Lenovo Outstanding Young Scientists Program (LOYS).

6. REFERENCES

[1] J. Bai, Y. Wu, J. Zhang, and F. Chen. Subset based deep learning for rgb-d object recognition. *Neurocomputing*, 165:280 – 292, 2015.

[2] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *Computer Vision & Image Understanding*, 110(3):404–417, 2006.

[3] M. Blum, J. T. Springenberg, J. Wulfing, and M. Riedmiller. A learned feature descriptor for object recognition in rgb-d data. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1298–1303, May 2012.

[4] L. Bo, X. Ren, and D. Fox. Depth kernel descriptors for object recognition. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 821–826, Sept 2011.

[5] L. Bo, X. Ren, and D. Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2115–2123. Curran Associates, Inc., 2011.

[6] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[7] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. *Learning Rich Features from RGB-D Images for Object Detection and Segmentation*. Springer International Publishing, 2014.

[8] I. H. Jhuo, S. Gao, L. Zhuang, D. T. Lee, and Y. Ma. *Unsupervised Feature Learning for RGB-D Image Classification*. Springer International Publishing, 2014.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[11] C. Y. Liou, W. C. Cheng, J. W. Liou, and D. R. Liou. Autoencoder for words. *Neurocomputing*, 139(139):84–96, 2014.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[13] M. Lukosevicius and H. Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

[14] X. Lv, S.-Q. Jiang, L. Herranz, and S. Wang. Rgb-d hand-held object recognition based on heterogeneous feature fusion. *Journal of Computer Science and Technology*, 30(2):340–352, 2015.

[15] X. Lv, X. Liu, X. Li, X. Li, S. Jiang, and Z. He. Modality-specific and hierarchical feature learning for rgb-d hand-held object recognition. *Multimedia Tools and Applications*, pages 1–18, 2016.

[16] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng. Convolutional-recursive deep learning for 3d object classification. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 656–664. Curran Associates, Inc., 2012.

[17] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *2009 IEEE 12th International Conference on Computer Vision*, pages 32–39, Sept 2009.