One-shot Learning for RGB-D Hand-Held Object Recognition

Jian Sun
Shandong University of Science and Technology
Qingdao, China
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
jian.sun@vipl.ict.ac.cn

Yaohui Zhu
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
yaohui.zhu@vipl.ict.ac.cn

Shuqiang Jiang
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
sqjiang@ict.ac.cn

ABSTRACT
With the advance of computer technology and smart device, many applications, such as face recognition and object recognition, have been developed to facilitate human-computer interaction (HCI) efficiently. In this respect, the hand-held object recognition plays an important role in HCI. It can be used not only to help computer understand user's intentions but also to meet user's requirements. In recent years, the appearance of convolutional neural networks (CNNs) greatly enhances the performance of object recognition and this technology has been applied to hand-held object recognition in some works. However, these supervised learning models need large number of labelled data and many iterations to train their large number of parameters. This is a huge challenge for HCI, because HCI need to deal with in-time and it is difficult to collect enough labeled data. Especially when a new category need to be learnt, it will spend a lot of time to update the model. In this work, we adopt the one-shot learning method to solve this problem. This method does not need to update the model when a new category need to be learnt. Moreover, depth image is robust to light and color variation. We fuse depth image information to harness the complementary relationship between the two modalities to improve the performance of hand-held object recognition. Experimental results on our handheld object dataset demonstrate that our method for hand-held object recognition achieves an improvement of performance.

CCS CONCEPTS
• Computing methodologies → Object recognition;

KEYWORDS
hand-held object recognition, RGB-D, Multimodal information fusion, one-shot learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICIMCS '18, August 17-19, Nanjing, China
© 2018 ACM.
ACM ISBN 978-1-4503-6520-8/18/08...
https://doi.org/10.1145/3240876.3240920

1 INTRODUCTION
Recently, with the development of computer technology, smart devices, such as smart robots, have become more and more common in people’s daily lives. Human-computer interaction (HCI) becomes an essential part of our life. The traditional way of HCI through keyboard input or controller operation is inconvenient to most people, especially for the elderly and children, so people need a more friendly way of HCI. For human, hands play an important role in communicating. When we show an object to others, we usually take it in our hands to attract more attentions. Similar to this, in HCI, hand-held objects also play important role. It can be used to help computer understand user’s intentions or user’s requirements. So if a computer can recognize what a person is holding, HCI will be more convenient for people. In this case, the study of hand-held object recognition is of great significance.

As we all know, the widely used technology in HCI is object recognition. Object recognition is a classical task in the field of computer vision and multimedia. But, in the real world, due to the
influence of various factors such as illumination, occlusion, deformation, etc, the task still face enormous challenges. In the past few years, various approaches have been proposed for object recognition and they try to find more robust hand-craft feature representation such as SIFT [9], SURF [1], etc, but those hand-craft features look the bottleneck, the performance of this task is hard to obtain big improvement. In recent years, deep learning technology regains public recognition and the learning-based features become the mainstream method. In the year of 2012, Krizhevsky et al proposed a classic Convolutional Neural Network (CNN) architecture which is named AlexNet [7] and show significant improvements upon previous methods on the object recognition task. Since then, CNN model become the most commonly used deep learning method for object recognition and other tasks such as scene recognition [18] [6].

Although deep learning methods have achieved great success in most computer vision tasks. However, these models use the supervised learning method which needs large number of labeled data and many iterations to train their huge parameters. This annotation of labeled data severely limits the flexibility of deep learning model to new classes, and more fundamentally limits their applicability to deal with newly emerging objects or rare categories where numerous annotated images may simply never exist. On the other hand, humans are very good at recognizing objects with very little or none direct supervision. Motivated by the failure of deep learning methods to work well on one or few examples for per class, and inspired by the few-shot learning ability of humans, more and more researchers begin to pay attention to few-shot learning, and some works has been proposed, such as [3] [15]. In this paper, we use few-shot learning in hand-held object recognition.

When people interact with a computer, the camera has the ability of capturing object images in real-time. In addition, with the recent advent of commodity depth cameras, an increasing amount of visual data not only contains color but also depth information. It is expected that the additional depth information will lead to improve object recognition performance due to the robustness of depth information to light and color variation [20].

In this work, we propose a new method to process the problem of hand-held object recognition which is based on one-shot learning. One-shot learning is well suited to handle this problem in hand-held object recognition, because it just need few examples to classify objects and once the model is trained, the model can classify images of new classes by computing relation scores between query images and the few examples of each new class without updating the model. Moreover, we fuse depth image information to harness the complementary relationship between the two modalities to improve the performance of hand-held object recognition. We evaluate the performance in two different levels: category level and instance level.

2 RELATED WORKS

2.1 RGB-D object recognition

Object recognition is a classical task in computer vision. Traditionally, only using RGB images to recognize objects is a challenging task as real-world scenes are often cluttered and have variable illumination. With the recent advent of commodity depth cameras, an increasing amount of visual data not only contains color but also depth measurements. It is expected that the additional depth information will lead to improved object recognition performance due to the robustness of depth measurements to light and color variation. It has been shown that depth image as a kind of supplementary information has a great help to object recognition. Yanhua Cheng et al [2] use depth image correcting errors in object classification in indoor semantic segmentation of images. Guillermo et al [5] use RGB-D videos to study the hand poses recognize with 3D objects. Charles et al [12] study 3D object detection from RGB-D data in both indoor and outdoor scenes and they directly operate on raw point clouds by popping up RGB-D scans. All these research results prove that depth image can effectively improve performance in object recognition.

Previous RGB-D object recognition models fuse RGB and depth information by three main modes: serializing RGB and depth as undifferentiated 4-channel input which can be called input fusion, combining handcrafted RGB and depth saliency features which can be called feature fusion, or performing unimodal predictions separately and then make joint decisions which can be called result fusion. In this work we adopt the first and second method to fusion the two modality informations.

2.2 Hand-held object recognition

The study of hand-held object recognition can be divided into first-person interface and second-person interface. Most previous works focused on first-person interface [13] [21] [8]. Ren et al. [13] 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. Xu et al. [21] is driven by a headpose calculation and laser pointer guidance to estimate the region of interest for the hand-held objects and the object at distance. They compute the region of interest to recognize the hand-held object with SIFT [9]. Literature [8] focus on continuous learning of hand-held object recognition. They propose a dataset, CORo50, which is a collection of 50 domestic objects belonging to 10 categories, and they also provide a benchmark for this task using CNN.

Moreover, many works focus on second-person interface. The images of second-person interface containing user, object and background. The main idea of these works is using RGB-D devices (i.e., Kinect) to capture both RGB and depth information, they take advantage of skeletal information of user to locate the hand position and depth information to segment the object. Lv et al [10] [11] uses the fine-tuned AlexNet [7] to extract the features of RGB images and depth images, and extract the hand-craft features of the 3D point cloud, at last, they fuse three types features as the final feature representation and achieved a good result. But they don’t consider the issue of how to learn a new knowledge.

2.3 Few-shot learning

The study of few-shot learning of object recognition has been of interest for some time [3]. Recently, with the success of deep learning-based approaches in the data-rich object recognition task, there
has been a surge of interest in applying such deep learning approaches to the few-shot learning field.

The successful MAML method [4] use meta-learning to train a model on a variety of learning tasks, such that it can solve new learning tasks using only a small number of training samples. Oriol et al proposed a metric learning based method [19] for one-shot learning. They employ ideas from metric learning based on deep neural features and from recent advances that augment neural networks with external memories. Their framework learns a network that maps a small labeled support set and an unlabeled example to its label, obviating the need for fine-tuning to adapt to new class types. Flood et al [22] improved the method in [19] they change the distance calculation from the fixed way, such as the Euclidean distance, to a learnable way, that can be learned by a neural network.

3 OUR APPROACH

In this work, we use the current popular one-shot learning method to explore one-shot hand-held object recognition. Further more, we propose a method based on multi-modal information fusion to enhance recognition performance. Our main work can be divided into two key parts: data collection and object recognition.

3.1 Data Collection

In this part, we apply the region growing algorithm [10] in depth image to remove background and segment out the object at the same time. This algorithm is a fundamental region-based image segmentation method. We use the location of user’s hand as the seed point and the criterion $f(x_i, y_i)$ is defined as equation (1). If $f(x_i, y_i)$ is 1, the pixel located in $(x_i, y_i)$ is divided into object, otherwise, it belongs to background.

$$f(x_i, y_i) = \begin{cases} 
1 & \text{abs}(D(x_i, y_i) - d_{mean}) \leq T \\
0 & \text{abs}(D(x_i, y_i) - d_{mean}) > T 
\end{cases}$$

(1)

The $D(x_i, y_i)$ indicates the pixel value in depth image at location $(x_i, y_i)$. $T$ is an empirical threshold and $d_{mean}$ is the average depth of the 8-connected neighborhood of the seed point. After initialization, the method keeps examining the 8-connected neighborhood of seed points. If any pixel meets the criterion, then treat it as new seed point. This process is iterating until there is no change in two successive iterative stages. The detail process is shown in Algorithm 1.

3.2 One-shot Hand-held Object Recognition

Many effective one-shot learning methods have been proposed, and all of these methods use a meta-learning method to implement learning to learn in order to achieve one-shot learning. Among these methods, the recently proposed Relation Network [22] is simple and efficient, and it can be trained in the way of end-to-end. The Relation Network is a CNN architecture which consists of two modules, encoder module and relation module. The encoder module take the images of support set and target set as input, and then encode them into feature maps. The feature maps of target set image is concatenated to the corresponding feature maps of each image in the support set. All the concatenated feature maps are sent into the relation module to calculate a similarity score and take the highest score as the predict result.

<table>
<thead>
<tr>
<th>Algorithm 1 region growing algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> depth image $d_x$, the seed point (position of hand) $x_h$, threshold $T$</td>
</tr>
<tr>
<td><strong>Output:</strong> the set of object mask: $S_x$</td>
</tr>
<tr>
<td><strong>Others:</strong> queue of position $Q$, binary image $L_x$, eight-neighbor pixels $N_x$</td>
</tr>
<tr>
<td>1. for all pixels $x$, $S_x = 0$, $L_x = 0$</td>
</tr>
<tr>
<td>2. $Q.push(x_h)$</td>
</tr>
<tr>
<td>3. for all $z$ in $N_x$, $d_h = (d + \sum z d_z)/9$</td>
</tr>
<tr>
<td>4. while $Q$ is not empty:</td>
</tr>
<tr>
<td>5. $q = Q.pop()$</td>
</tr>
<tr>
<td>6. if $L_q = 1$ and $</td>
</tr>
<tr>
<td>7. $S_q = 1$</td>
</tr>
<tr>
<td>8. for every point $z$ in $N_x h$</td>
</tr>
<tr>
<td>9. if $L_z == 1$</td>
</tr>
<tr>
<td>10. $Q.push(z)$</td>
</tr>
<tr>
<td>11. $L_q = 1$</td>
</tr>
<tr>
<td>12. return $S_x$</td>
</tr>
</tbody>
</table>

RelationNetwork use the contrast of two images to achieve metalearning, the model inspired by humans, that people recognize images by comparing the features between images and images to achieve recognition, that is, one-shot learning; of course, people can also compare images and the description of one class to achieve the zero-shot learning. Because our visual cells can automatically extract image features, such as outlines, light intensity, etc. and then compare our past experience to identify images. The two modules in RelationNetwork are just like the process of humans recognizing objects when performing object recognition. The first module extract image features and the second module do object recognition by comparison two features of two images.

The model is trained with the mean square error loss function as shown in (2). In this formula, $\varphi$ and $\omega$ indicate parameters of encode module and relation module, $r_{i,j}$ is the last result of relation module.

$$\varphi, \omega = \arg\min_{\varphi, \omega} \sum_{i=1}^{m} \sum_{j=1}^{n} (r_{i,j})^{-1}(y_i == y_j)$$

(2)

For only using RGB images, $r_{i,j}$ can be calculated by the following formula (3). $g_{abl}()$ represent the relation module, $f_{abl}()$ represent-the encoder module, $C(\cdot)$ means concatenate two features.

$$r_{i,j} = g_{abl}(C(f_{abl}(x_i), f_{abl}(x_j)))$$

(3)

For the multi-modal information fusion, $r_{i,j}$ is calculated by the formula (4) or formula (5). In (4) we take the concatenated RGB and depth images as the input of RelationNetwork, and in (5) we pass the encoder module and directly take the concatenated VGG-16 feature maps as the input of relation module. $v$ means the VGG-16 feature maps, $rgb$ means RGB image and $dep$ means depth image.

$$r_{i,j} = g_{abl}(C(f_{abl}(C(rgb_i, dep_i)), f_{abl}(C(rgb_j, dep_j))))$$

(4)

$$r_{i,j} = g_{abl}(C(C(v(rgb_i), v(dep_i)), C(v(rgb_j), v(dep_j))))$$

(5)
4 EXPERIMENTS

4.1 Dataset
In order to evaluate our method, we collected a dataset which is named HOD-40 (Hand-Held Object Dataset). The dataset contains 32000 video frames including 40 categories and every category has 4 different instances. Partial categories and instances of this dataset are shown in Figure 2. For each category, we let 2 people collect in 2 scenes, and we collect 200 frames from each scene with the Kinect mounted at chest height. Each frame captures the visual image and the corresponding depth map.

4.2 Experimental Setup
Our all experiments are performed on the HOD-40 dataset. Not like the general classification task, there are different label space in the one-shot learning problem between the train and the test phases. We randomly select 30 classes as training set and the remaining 10 classes as the test set. Just as Figure 1 shows, the support set contains 1 labeled examples for each of 5 unique classes, the target problem is called 5-way 1-shot learning. We randomly choose 5 classes from training set as support set and randomly choose a class from the 5 classes as target set. We choose one image from each class in support set and 15 images in target set and take these images as one episodes.

In the training phase, we randomly generated 30000 episodes for one epoch, and total run 10 epoches. We use the common stochastic gradient descent to optimize the model. The initial learning rate is set to 1e-4 and decrease by half every 4 epoches. In the test phase, we randomly generated 2000 episodes to get the average accuracy. The input images are all resized to 84 x 84.

We evaluate the performance in two different levels, just as proposed in the literature [20], category level and instance level. In instance level, the images of target set come from the same instance with correspondent class in support set. On the contrary, in category level, the images of target set only come from the same category maybe not the same instance with correspondent class in support set. We train the model in category level but test the model in two levels. All experiments use the same parameter settings.

4.3 Results And Analysis

4.3.1 RGB Images. In this part we give the result of one-shot learning only using RGB images in two different levels, the accuracy of 10 epoches are shown in Figure 3. Obviously, from this result we can see the accuracy of instance level is higher than category level. This is easy to explain, because in instance level, the images in target set and support set come from the same instance, in contrast, the images in category level may come from different instance, they have higher inner-class difference than instance level.

From this result we can see that one-shot learning is suitable for the task of hand-held object recognition. First, the accuracy of one-shot learning in hand-held object recognition is not bad and acceptable. Second, in the process of one-shot learning the images in test phase are not seen during training phase, so when a new category needs to be learnt, we not only don’t need to collect a huge amount of data, but also no need to update the model.

4.3.2 RGB-D Images. We take the result of using only the RGB image as the baseline and explore the effect of depth information on recognition accuracy with one-shot learning. In the experiment of RGB and depth information fusion, we try two different methods.

\[
x_i' = (y_{\text{max}} - y_{\text{min}}) \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} + y_{\text{min}}
\]

The first fusion method is simply concatenate RGB image and it’s correspondingly depth image on channel dimension. RGB image and depth image are two different modality information. The two kind of information have different range in pixel values. So, simply concatenate this two images requires the network to automatically adjust the amplitude between the two modalities. This may leads to slowing down the weight learning procedure and influencing the final performances as well. Therefore, we first implement normalization on both depth and RGB images to restrict the values of them between 0 and 1 and then concatenate them. The normalization is shown in (6). The \(x_{\text{max}}\) and \(x_{\text{min}}\) are the maximum and minimum of all pixels in the same channel, \(y_{\text{max}}\) and \(y_{\text{min}}\) are the maximum and minimum of target variable range, \(x_i\) is the pixel value of image, \(x_i'\) is the corresponding value after normalization. After this concatenation, we take the 4 channels images as input to train Relation Network. For comparison we use the same parameters with the experiment that only use RGB images. The result is shown in Figure 4. The accuracy has been improved both in category level and on
The experimental results show that the features extracted from the different modes of RGB and depth have better expressive capabilities, and can more effectively fuse the information of two different modes of RGB and depth. The second information fusion method we take is at the feature level. The architecture is shown in Figure 1. We separately use the RGB image and depth image to fine tune the pre-trained VGG-16 model and then use the fine-tuned models to extract conv5_3 layer feature maps. Similar to the previous experiment, we also normalize the features to restrict them between 0 and 1. We concatenate the feature maps of RGB image and depth image and then put them directly into the relation module to learn the last scores. This is equivalent to using the VGG-16 network instead of the previous encoder module. Table 1 shows the accuracy of 3 different methods used in this article. The accuracy is the maximum of 10 epochs. The experimental results show that the features extracted from the deeper network have stronger feature representation capabilities and can effectively improve the accuracy of one-shot learning.

Table 1: Accuracy of different method in two levels

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>category level</th>
<th>instance level</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB</td>
<td>0.747</td>
<td>0.838</td>
<td></td>
</tr>
<tr>
<td>RGB-D(image)</td>
<td>0.762</td>
<td>0.849</td>
<td></td>
</tr>
<tr>
<td>RGB-D(feature)</td>
<td>0.825</td>
<td>0.860</td>
<td></td>
</tr>
</tbody>
</table>

We do this by taking into account that the encoding module in the original RelationNetwork has only four convolutional layers and this has weaker encoding capabilities for images. However, after adding a depth image, the entire input image becomes more complex, so a deeper network is needed. With the deeper encoder module, the features thus obtained have better expressive capabilities, and can also more effectively fuse the information of two different modes of RGB and depth.

5 CONCLUSION

In this paper, we present a new method for hand-held object recognition which is based on one-shot learning. We use Relation Network to train our classification model, and the Relation network learns an embedding and a deep non-linear distance metric for comparing query and sample items. Based on this, we fuse depth image information as an auxiliary information and effectively improving the accuracy of hand-held object recognition. To evaluate our method, we collect a dataset contains 40 categories. All our experiments are done on this data set, and all the results show that our method is very suitable for hand-held object recognition.

6 ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61532018, in part by the Lenovo Outstanding Young Scientists Program, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals.

REFERENCES


