Real-time hand grasp recognition using weakly supervised two-stage convolutional neural networks for understanding manipulation actions

Ji Woong Kim, Sujeong You, Sang Hoon Ji, and Hong Seok Kim
Korea Institute of Industrial Technology
Republic of Korea
kjw6139@gmail.com, {sjyou21, robot91, hskim}@kitech.re.kr

Abstract

Understanding human hand usage is one of the richest information source to recognize human manipulation actions. Since humans use various tools during actions, grasp recognition gives important cues to figure out humans' intention and tasks. Earlier studies analyzed grasps with positions of hand joints by attaching sensors, but since these types of sensors prevent humans from naturally conducting actions, visual approaches have been focused in recent years. Convolutional neural networks require a vast annotated dataset, but, to our knowledge, no human grasping dataset includes ground truth of hand regions. In this paper, we propose a grasp recognition method only with image-level labels by the weakly supervised learning framework. In addition, we split the grasp recognition process into two stages that are hand localization and grasp classification so as to speed up. Experimental results demonstrate that the proposed method outperforms existing methods and can perform in real-time.

1. Introduction

Human action recognition has many applications in robot learning from demonstration and human-robot-interaction. Usually, humans use various types of hand grasp according to intention of the action and the kind of tools. Hence, human grasp recognition is a crucial key to understand human manipulation actions [1], and various methods [2] of the grasp recognition have been explored based on the grasp taxonomy developed by Feix *et al*. [3]. However, there are several properties of grasping such as resemblance among different functional grasp types and occlusion of hands by grasped objects that make the grasp recognition difficult. Thus, in this work, the CNN framework is utilized to overcome these challenging aspects. However, a large well-annotated dataset is an important prerequisite to use the CNN framework, and to our knowledge, no human grasping dataset with ground truth of hand region has existed so far. Thus, we propose two-stage convolutional neural networks, where one CNN was trained as a hand localization network with weakly supervised learning framework [4] and the other CNN was trained as a grasp classification network with patches around detected hand locations. The experimental results demonstrate that our method outperforms the existing methods and can be performed at more than 60 fps which is enough for real-time processing.

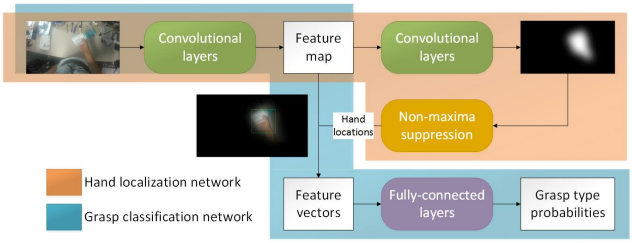

Figure 1. The network architecture and the data flow of our approach.

2. Our approach

2.1. Architecture

The representative CNN architectures, such as AlexNet [5] and VGG-16 network [6], are comprised of two phases. First, the CNNs extract a feature map from an input image using convolutional layers, and then the class of an object in the image is determined by fully connected layers with the feature map. Since human grasping datasets do not include ground truth bounding boxes of hands, the weakly supervised learning framework [4] is utilized to coarsely predict locations of grasp instances. For the sake of that, every fully connected layer is converted to the equivalent convolutional layer, and then the network produces a probability map of each grasp type. However, because hand grasp images in the videos of object manipulation contain few hands in each image and grasp instances rarely overlap each other,

probability values peak exclusively at few locations. Hence, we change the fully convolutional network for grasp type recognition into a two-stage CNN so that classifying grasp type is performed only on the hand regions which are found by the weakly supervised learned hand localization network with non-maxima suppression.

Because the information of feature map of the grasp classification network will be enough for judging the presence of hands, we modify the architecture so that the two networks share the feature map. Furthermore, the output channels of the convolutional layers in the class-determination phase are set to 128. The overall network architecture of proposed method is shown in Fig. 1.

2.2. Training

Our network architecture was applied to the AlexNet and the VGG-16 network. The network training took two steps. First, the grasp classification network was initialized by the pre-trained model trained on the ILSVRC dataset and fine-tuned on the university of Tokyo(UT) grasp dataset [2]. Next, the feature extractor layers were fixed, and rest layers of the hand localization network were initialized randomly. Then, the hand localization network was fine-tuned on the hand training set composed of the same images with the grasp dataset but labeled by only two classes. In order to obtain hand grasp images, we tracked hand positions using the mean-shift tracking algorithm [7], cropped 600x600 patches around the positions, and resized them to 224x224. At test-time, we resized test input images so that the height is 600. Then, negative examples were created by randomly cropping a background patch per image. We created a training and a test set by dividing all images of each class at a ratio of 8 to 2, respectively.

3. Experimental results

We utilized Caffe [8] on a desktop equipped with NVIDIA TITAN X GPU, and all optimization processes were performed using the stochastic gradient descent with 0.001 of learning rate and 0.9 of momentum. The performance of each classification method is represented in terms of mean values of average precision, accuracy, and F1-measure of all classes and summarized in Table 1. Compared to the baseline method of Cai *et al*. [2] whose mean F1-measure is 89%, the results demonstrate that our proposed method surpasses the baseline. Additionally, due to the more discriminative power of the VGG-16, the classification performance of our VGG-16-based network is better than that of the AlexNet-based network. Our proposed networks took under 100 milliseconds per image during whole process. Especially, our AlexNet-based network achieved 62.76 fps can be used for grasp recognition in real-time.

Table 1. Summary of the classification results

network type	AlexNet-based	VGG-16-based
mAP (%)	93.16	95.46
mACC (%)	99.86	99.96
mF1 (%)	98.51	99.57
runtime (ms)	15.93	88.80

4. Conclusion

We propose a method for recognizing hand grasps automatically by using a two-stage CNN architecture. Our method improves recognition performance compared with existing methods, and it is verified that our AlexNet-based network can be used in real-time applications.

This work is actually an on-going work, thus there are many future works. More experiments should be carried out on the various grasping datasets. In addition, we will investigate the ways of incorporating temporal information and extending to recognition of manipulation actions.

5. Acknowledgement

This work was supported by the Industrial Strategic Technology Development Program (10048320, 10067414) funded by the Ministry of Trade, Industry & Energy (MOTIE, Republic of Korea)

References

[1] K. Gupta, D. Burschka, and A. Bhavsar. Effectiveness of grasp attributes and motion-constraints for fine-grained recognition of object manipulation actions. In *Proc. CVPR*, pages 1232–1239, 2016.

[2] M. Cai, K. Kitani, and Y. Sato. A scalable approach for understanding the visual structures of hand grasps. In *Proc. ICRA*, pages 1360–1366, 2015.

[3] T. Feix, J. Romero, H. Schmiedmayer, A. Dollar, and D. Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on Human-Machine Systems*, 46(1):66–77, 2016.

[4] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *Proc. CVPR*, pages 685–694, 2015.

[5] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, pages 1097–1105, 2012.

[6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015.

[7] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. CVPR*, pages 142–149, 2000.

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACM Multimedia*, pages 675–678, 2014.