Estimation of Object Functions Using Convolutional Neural Network

Yosuke Kitano
Graduate School of System Informatics,
Kobe University,
Nada, Kobe 657-8501,Japan
Email: kitano@me.cs.scitec.kobe-u.ac.jp

Tetsuya Takiguchi
Organization of Advanced Science
and Technology, Kobe University,
Nada, Kobe 657-8501, Japan
Email:takigu@kobe-u.ac.jp

Yasuo Ariki
Organization of Advanced Science
and Technology, Kobe University,
Nada, Kobe 657-8501, Japan
Email:ariki@kobe-u.ac.jp

***Abstract*—In recent years, a tremendous research effort has been made in the area of generic object recognition. However, both an object's name and function are important for robots to comprehend objects. Object functions refer to "the purpose that something has or the job that someone or something does". Various elements (e.g., the physical information, material, appearance and human interaction) independently or mutually form object functions. There are many researches on object functions using human-object interaction, while there are few using appearance. However, it can be believed that object functions may be formed by appearance. In this paper, we propose a new method to estimate object functions from appearance on images. Our approach is to estimate object functions using convolutional neural network(CNN), which has ability to learn rich mid-level features. In our method, we add adaptation layers for object function estimation formed by full-connection to the CNN which is pre-trained on the ImageNet2013 with 1000 object classes. Experimental results show that the classification rates of two of three functions of objects such as "cuttable" and "movable" are over** 80% **and that the appearance is closely related to object functions.**

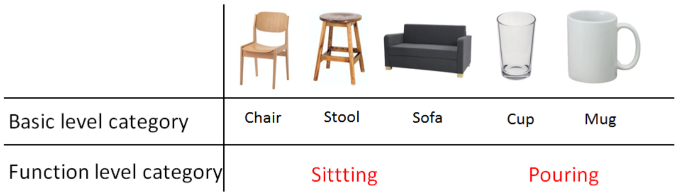

Fig. 1: Basic level categories vs. function level categories.

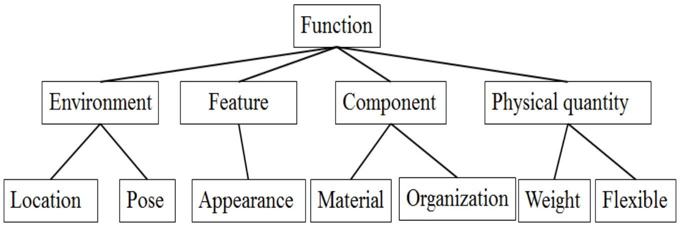

Fig. 2: Function-based ontology

I. Introduction

Object recognition means computer recognition of objects in a real world in terms of their generic names. It is one of the most challenging tasks in the field of computer vision. "Generic category of objects"[1] defines generic names as the basic level categories such as "chair" and "cup" in the area of object recognition. A practical example of generic object recognition is that household robots identify objects specified by human voice[2], [3]. For example, when an user asks the robot to bring the pen, it identifies and brings the pen if it knows the pen in advance.

However, there is a question if it is enough for robots to simply learn the object names and images. Since objects, the artifact we daily use, are made with their purposes, it is possible to regard objects as the means to accomplish the purpose.

In the above example, it can be thought that "we use the pen (means) to accomplish the purpose of writing (function)". Therefore, for robots to identify the object, both the object name such as "pen" and the function such as "allowing us to write" should be recognized. If the robot can estimate the object functions, even in the case there is no pen in the circumstances, the robot can bring the substitution such as "a writing brush" for us to write.

The above mentioned example, "bring me a pen" is the case where human specifies the object name and the robot knows the object but can't find the object so that it managed to find the substitution of the pen. However, even the robot does not know the object name, we want the robot to find the object where can be used as a writing tool.

We show the example of basic level category and function level category of objects in Fig. 1. In this paper, recognizing objects in the basic level category is defined as generic object recognition and recognizing objects in the function level category as function estimation. Today, a tremendous research effort has been made in the area of generic object recognition. In contrast to it, there is a few researches on function estimation, because functional class has a wide variety in the appearance and attributes forming the function. However, function estimation has begun to be focused on because many kinds of sensors are developed and it has become easy to observe the attributes possessed by the objects.

Fig. 2 shows the function-based ontology, which can be

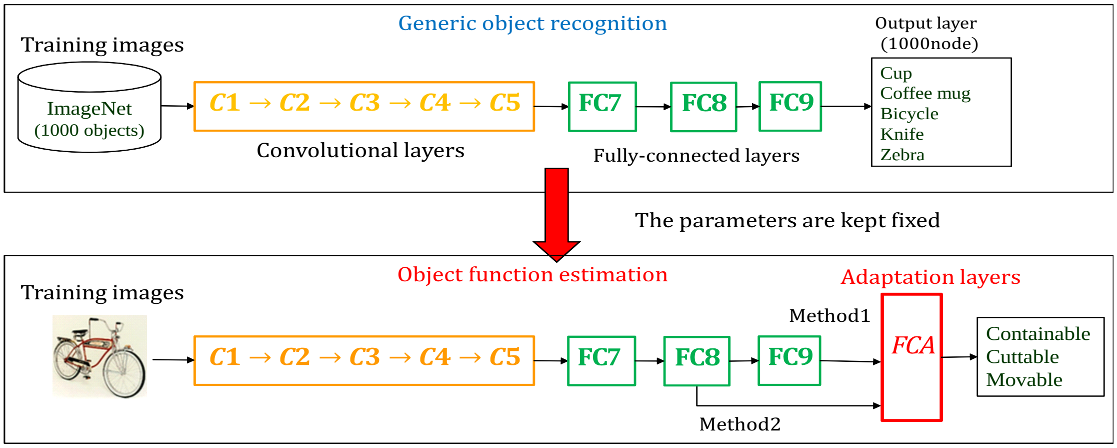

Fig. 3: Overview of proposed method

induced from the idea of Eric Wang[4]. It is assumed that various elements (e.g., the physical quantity, material, appearance and human interaction, environment) independently or mutually form object functions.

In this work, it is presumed that object functions are closely related to the appearance. However, the relationship between the object functions and the appearance is ambiguous. To address this problem, we estimate object functions using CNN which shows the state-of-the-art performances in various tasks such as object category recognition, handwritten character recognition and scene recognition. CNN is a classifier with very large number of parameters which must be learned from tremendous training images. In addition, CNN which is pre-trained on the dataset can be seen as an extractor of mid-level image representation. In this work, taking advantage of this idea, we estimate object functions by adding adaptation layers for object function estimation to the CNN which is pre-trained to recognize the object in images. This adaptation layers are formed as full connected layers. In the experiment, we test the unknown object images to evaluate whether we can train the network of the object functions on the object recognition network.

By adding adaptation layers on the object recognition CNN, the network system can find the object with the specified function by human even it does not know the object itself.

The rest of this paper is organized as follows: In Section 2, related works are described and our method is proposed in Section 3. In Section 4, the experimental data is evaluated, and the final section is devoted to our conclusions and future work.

II. Related Work

First, we distinguish function from affordance. It says in the dictionary that function refers to “the purpose that something has or the job that someone or something does”. American psychologist James.J.Gibson coined the term affordance[6]. Gibson and his colleagues argue that affordance refers to the quality of objects or environment that allows humans to perform some actions[7]. In the field of computer vision, research about affordance is popular. The interpretation of affordance is different a little among them. According to [8], [9], they define affordance as the relationship between robotics hand and objects, while according to [10], they define affordance as functionality in human action. As mentioned above, it is assumed that function is more comprehensive expression than affordance, and affordance is the function which depends on environment or human action.

There are a lot of researches about affordance, whose task or environment is limited. In [11], [12], they set up the task that makes the robot search for the object where humans can sit. In [13], humans might interact with the same object in different ways, with only some typical interactions corresponding to object affordance. [10], [14] show that they represent objects in the kitchen directly in terms of affordance. They model correlation between all object-object and human-object interactions. However, the task or environment is so limited that the number of objects is too limited. Thus it can be thought that, for function estimation, specific object recognition is carried out with the functional label annotated in advance. In this work, we estimate the object functions without limiting the task or environment. If we estimate the object function using interaction between human and object, we have to limit the task or environment as mentioned above. Therefore we estimate the object functions from their appearance on the image containing the single object.

Convolutional Neural Network(CNN), proposed by LeCun et al.[15], has shown grate performances in various computer vision applications, such as hand written character recognition[16], facial analysis [17]. CNN consists of a pipeline of convolution and pooling operations followed by a multi-layer perceptron. They tightly couples local feature extraction, global model construction and classification in a single architecture where all parameters are learned conjointly using back-propagation. CNN has been shown that it has an ability to learn a richer and more discriminative feature

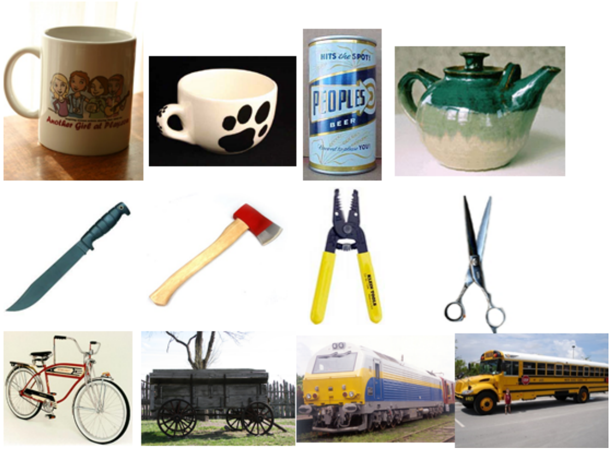

Fig. 4: Image examples in ImageNet

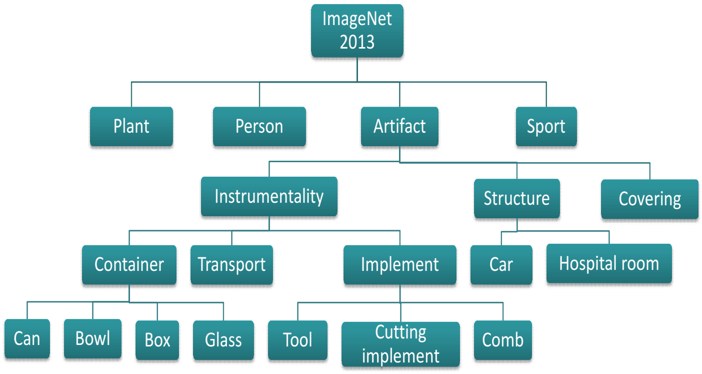

Fig. 5: Overview of WordNet

mapping than hand-crafted features such as HOG and SIFT across various vision tasks. In this work, we estimate object functions by adding the adaptation layers to the CNN which is pre-trained to recognize objects in images.

III. Function estimation using CNN

The architecture of CNN contains 60 millions of parameters. Therefore, a large amount of training images are needed to learn CNN. In recent work, the features extracted by a CNN trained on ImageNet are enough to achieve state-of-the-art results in image classification tasks. Based on the fact, we use CNN as a extractor of mid-level representation. In this work, we use CNN which is pre-trained to recognize objects in images. This training is done on the ImageNet2013 with 1000 object classes. The net work takes as input a square 221×221 pixel RGB images. This network is composed of six convolutional layer followed by three fully connected layers. Using shorthand notation, the full architecture of CNN is $C1(96, 7, 3)$-N-P-$C2(256, 7, 1)$-N-P-$C3(512, 3, 1)$-$C4(512, 3, 1)$-$C5(1024, 3, 1)$-$C6(1024, 3, 1)$-P-$FC7(4096)$-$FC8(4096)$-$FC9(1000)$, where $C(c, f, s)$ indicates a layer with c channels of $f \times f$ size applied with a stride s. $FC(n)$ is a fully-connected layer with n nodes. N is the normalization layer and P is the pooling layer. There is a detailed description of this network in [19] All convolutional layer and fully-connected layers use the rectified linear unit (ReLu) non-linear activation functions. An overview of our approach is shown in Fig. 3.

To achieve function estimation, we add the adaptation layers formed by full-connected layers to CNN which is trained to recognize object in images. The adaptation layers use the output vector of layer FC8 or FC9 as input. Here, it is considered that output vector of FC8 is related to semantic attribute and that of FC9 is label of generic object. In this paper, the function estimation is called the method 1 which takes the output vector of layer FC9 as the input of the adaptation layers. In the same way, the method 2 takes the output vector of FC 8 as the input. The parameters of CNN which is trained to recognize objects in images are kept fixed, then only adaptation layers are trained to estimate object functions. In training the specific object function, we collected positive images with the objects and negative images without the objects.

IV. Experiments

A. Dataset

In this experiment, we collected the images from ImageNet[18]. It is an image database formed based on the WordNet hierarchy, in which each node in the hierarchy corresponds to the synset. Here, synset is the group of a set of synonyms. The reason why we collect the images from ImageNet is that we can associate functions with synsets.

The task of function estimation is carried out for 3 classes (“movable”, “cuttable”, “containable”). We collected cup, kettle, can, pod, vessel and bin as “containable”. In the same way, bicycle, train, wagon, bus, sport car and scooter were collected as “movable” and knife, scissors, ax, punch, blade and plane as “cuttable” (see Fig. 4).

This is because the above three functions can be expressed by appearance. Fig. 5 shows the overview of WordNet. We collected the “containable” objects from “container” node in WordNet, the “cuttable” objects from “implement” node and the “movable” objects from “transport” node in WordNet. Here, “wagon” and “bicycle” are originally included in “container” node in WordNet, but we regard them as “movable” function because we usually regard them as ”movable” function objects rather than ”containable”. The number of training images and test images were about 8000 and 1000 images per function class respectively.

B. Experimental condition

In this experiment, we used the OverFeat[19]. Overfeat is the CNN which is pre-trained using 1,281,167 images in the CLS-LOC dataset of ILSVRC2013. The experiment was done in various number of layers and nodes with adaptation layer to train the object functions. In addition, we evaluate

TABLE I: Classification rates of Method 1.（%）

	The number of adaptation layers		
	1 layer	2 layers	3 layers
Containable	75.3	77.6	71.2
Cuttable	86.2	86.8	86.2
Movable	86.9	86.7	85.3

our proposed model using cross-validation. For instance, in calculating the classification rate of ”containable” function, we collected many images of ”cup” as test data, and the rest images without ”cup” as training data. This operation was done for each object which has ”containable” function. Then the classification rate for ”containable” function is attained by averaging the classification rate for each object.

C. Experimental result

TABLE I and TABLE II show the classification results by the method 1 and method 2 with the different number of the adaptation layers. In case where the number of adaptation layers is two, the average of classification result for function estimation by method 1 is highest and 83.7%. In case where the number of adaptation layers is three, the average of classification result for function estimation is highest and 84.9% by method 2. The average of classification result by method 2 is higher than that by method 1. It is considered that the feature of method 2 is richer mid-level representations than that of method 1. This is because the feature of method 2 is related to semantic attributes and that of method 1 is related to object class label.

However, we lack the understanding of why the method 1 and method 2 work so well. Therefore, we will have to analyze the configuration within CNN. For example, we will analyze the most important object in ILSVRC dataset within 1000 object classes for object function estimation in terms of parameters of adaptation layers.

V. Conclusion and Future Work

Various elements independently or mutually express the object function. We believe that function is closely related to the appearance, so we proposed the method that could estimate the object function using CNN. Function estimation of two of three classes such as ”cuttable” and ”movable” had over 80% accuracy in the experiments. Our experiments have shown that object functions could be formed by appearance. However, we lack understanding of why the classification rate of object function estimation is high. Therefore, we will analyze the reason experimentally.

In the future, the method of function estimation will be extended in two ways. Firstly, we model an input image using all activations in the network. We believe that object function can be predicted by a simple liner combination of CNN activation using sparse modeling.

Secondly, we attempt to employ part-based CNN. In [20], we argued that object function is closely related to the object part.

TABLE II: Classification rates of Method 2.（%）

	The number of adaptation layers		
	1 layer	2 layers	3 layers
Containable	43.2	82.2	82.9
Cuttable	32.2	85.5	85.8
Movable	79.8	86.5	85.9

References

[1] Rosch, Eleanor, et al. ”Basic objects in natural categories.” Cognitive psychology 8.3 (1976): 382-439.

[2] Nishimura, Hitoshi, et al. ”Object Recognition by Integrated Information Using Web Images.” Pattern Recognition (ACPR), 2013 2nd IAPR Asian Conference on. IEEE, 2013.

[3] Nishimura, Hitoshi, et al. ”Selection of an Object Requested by Speech Based on Generic Object Recognition.” Proceedings of the 2014 Workshop on Multimodal, Multi-Party, Real-World Human-Robot Interaction. ACM, 2014.

[4] Wang, Eric, Yong Se Kim, and Sung Ah Kim. ”An object ontology using form-function reasoning to support robot context understanding.” Computer-Aided Design and Applications 2.6 (2005): 815-824.

[5] Felzenszwalb, Pedro F., et al. ”Object detection with discriminatively trained part-based models.” Pattern Analysis and Machine Intelligence, IEEE Transactions on 32.9 (2010): 1627-1645.

[6] Gibson, James J. The ecological approach to visual perception. Psychology Press, 2013.

[7] Gibson, Eleanor J. ”The concept of affordances in development: The renascence of functionalism.” The concept of development: The Minnesota symposia on child psychology. Vol. 15. Hillsdale, NJ: Lawrence Erlbaum Associates Inc, 1982.

[8] Saxena, Ashutosh, Justin Driemeyer, and Andrew Y. Ng. ”Robotic grasping of novel objects using vision.” The International Journal of Robotics Research 27.2 (2008): 157-173.

[9] Stark, Michael, et al. ”Functional object class detection based on learned affordance cues.” Computer Vision Systems. Springer Berlin Heidelberg, 2008. 435-444.

[10] Pieropan, Alessandro, Carl Henrik Ek, and Hedvig Kjellstrom. ”Functional object descriptors for human activity modeling.” Robotics and Automation (ICRA), 2013 IEEE International Conference on. IEEE, 2013.

[11] Jiang, Yun, Marcus Lim, and Ashutosh Saxena. ”Learning object arrangements in 3d scenes using human context.” arXiv preprint arXiv:1206.6462 (2012).

[12] Grabner, Helmut, Juergen Gall, and Luc Van Gool. ”What makes a chair a chair?.” Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[13] Yao, Bangpeng, Jiayuan Ma, and Li Fei-Fei. ”Discovering object functionality.” Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, 2013.

[14] Pieropan, Alessandro, Carl Henrik Ek, and Hedvig Kjellstr?m. ”Recognizing Object Affordances in Terms of Spatio-Temporal Object-Object Relationships.”

[15] LeCun, Yann, et al. ”Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86.11 (1998): 2278-2324.

[16] Garcia, Christophe, and Manolis Delakis. ”Convolutional face finder: A neural architecture for fast and robust face detection.” Pattern Analysis and Machine Intelligence, IEEE Transactions on 26.11 (2004): 1408-1423.

[17] Delakis, Manolis, and Christophe Garcia. ”text Detection with Convolutional Neural Networks.” VISAPP (2). 2008.

[18] Deng, Jia, et al. ”Imagenet: A large-scale hierarchical image database.” Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[19] Sermanet, Pierre, et al. ”Overfeat: Integrated recognition, localization and detection using convolutional networks.” arXiv preprint arXiv:1312.6229 (2013).

[20] Kitano, Yosuke, Tetsuya Takiguchi, and Yasuo Ariki. ”Estimation of object functions using deformable part model.” Frontiers of Computer Vision (FCV), 2015 21st Korea-Japan Joint Workshop on. IEEE, 2015.