Estimation of Object Functions Focusing on Feature of Object Parts

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

Tetsuya Takiguchi
Graduate School of System Informatics,
Kobe University,
Nada, Kobe 657-8501, Japan
Email:takigu@kobe-u.ac.jp

Yasuo Ariki
Graduate School of System Informatics,
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 the 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 under the assumption that object parts contribute to estimating function. The rationale of the assumption is that when humans estimate function of unknown object, they focus on not only whole the object but each part of the object. In our previous method, object function was estimated by using mid-level feature of CNN which was pre-trained on the ImageNet2013 with 1000 object classes. In our proposed method, in addition to the mid-level feature, we use feature of object parts extracted from Deformable Parts Model(DPM) and Convolutive Bottleneck Network(CBN). Experimental results show that the classification rate of five functions is improved by 5.3% compared with the previous method.**

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

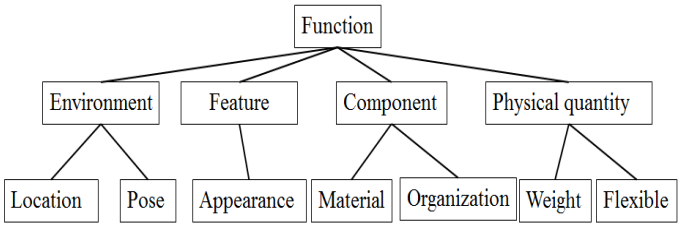

Fig. 2: Function-based ontology

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 when the robot does not know the object name, we want the robot to find the object which 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

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. In addition, we hypothesize that object parts contribute to estimating function. Because when human estimate function of unknown object, they focus on not only whole the object but each part of the object. Therefore, to estimate function, we use two features, namely, feature of object parts and feature of whole the object. We extract object parts using DPM[5], and then extract bottleneck feature of each parts using CBN[6]. In addition, we extract feature of whole the object by using CNN. CNN which is pre-trained on the dataset can be seen as an extractor of mid-level image representation. We merge their features, namely, bottleneck feature of each part and mid-level feature of CNN, and then estimate the function by using it as an input to MLP(Multi-Layer Perceptron). We execute experiment on the unknown object images to evaluate whether we can train the network of the object functions.

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[7]. Gibson and his colleagues argue that affordance refers to the quality of objects or environment that allows humans to perform some actions[8]. In the field of computer vision, research about affordance is popular. The interpretation of affordance is different a little among them. According to [9], [10], they define affordance as the relationship between robotics hand and objects, while according to [11], 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 [12], [13], they set up the task that makes the robot search for the object where humans can sit. In [14], humans might interact with the same object in different ways, with only some typical interactions corresponding to object affordance. [11], [15] 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 our 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. In [16], to estimate object function, CNN was pre-trained on the ImageNet 2013 with 1000 object classes and then used as an extractor of mid-level representation.

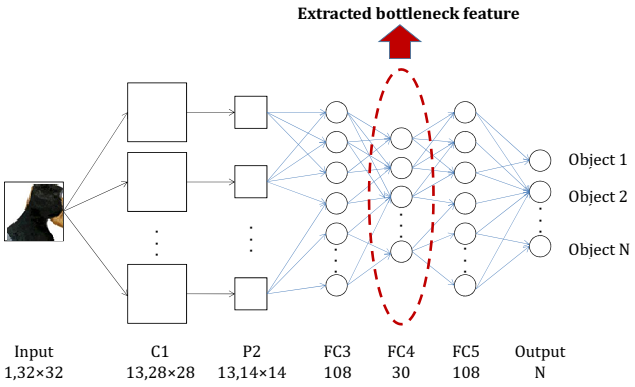

Fig. 3: Architecture of CBN

In addition to this mid-level feature, we also use feature of object parts extracted by using DPM. DPM represents objects by a lower-resolution root filter and a set of higher-resolution part filters arranged in a flexible spatial configuration. The part locations are treated as latent information. All the parameters of DPM are learned by LSVM(Latent SVM) which deals with the latent information. The LSVM learning procedure acquires part appearance and layout parameters by alternately performing the assignments to latent variables given the model parameters and re-optimizing the model parameters given the latent variable assignments. This system can detect objects over a wide range of scales and poses. It is unclear whether parts detected by DPM are related to function or not, but if function classification rate is higher than the previous one which does not employ the object parts, we can inductively show that object parts contribute to estimating function.

III. Function estimation using Feature of Parts

Firstly, we crop object parts by using DPM. DPM is trained against each object image of dataset. Using trained DPM, object detection is performed against each trained object image. Then, we crop object parts detected by part filters.

Secondly, bottleneck feature is extracted from each parts as shown in Fig.3. We trained CBN[6], taking as input a square of 32×32 pixel gray scale images of parts cropped by DPM, and as output object label. Using shorthand notation, the full architecture of CBN is $C1(13,5,1)$-$P2$-$FC3(108)$-$FC4(30)$-$FC5(108)$-$FC6(N)$, where $C(c,f,s)$ indicates a layer with c channels of $f \times f$ size filter applied with a stride s. $FC(n)$ is a fully-connected layer with n nodes. P is a pooling layer which pools spatially in non-overlapping 2×2 regions. After training CBN, we extract bottleneck feature($FC4$). Bottleneck feature represents much information in less nodes, so it has important feature associating input with output. In this case, bottleneck

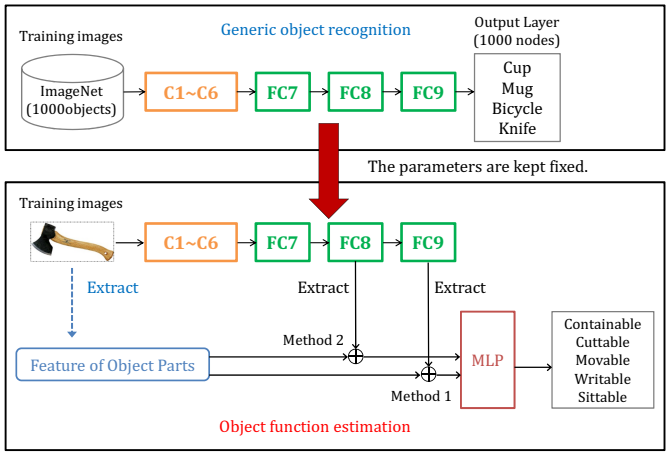

Fig. 4: Overview of proposed method

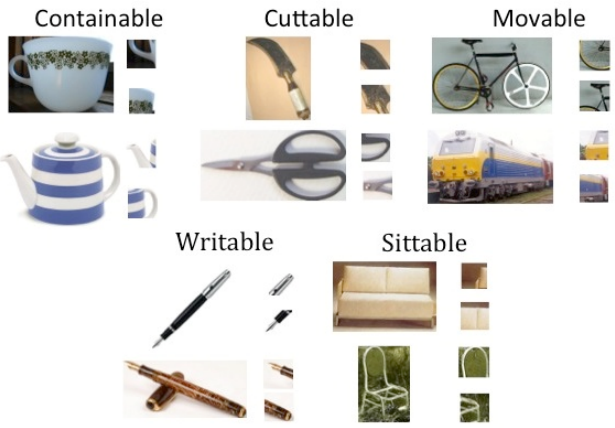

Fig. 5: Image examples of object and parts

feature represents potential relationship between object name and object parts.

Finally, we estimate object function using bottleneck feature of object parts as shown in Fig.4. In our previous method[16], we estimated object function using mid-level feature of CNN which was pre-trained on the ImageNet2013 with 1000 object classes(shown in Fig.4 up). To achieve function estimation, MLP was added as full-connected layers to CNN which was trained to recognize object in images(shown in Fig.4 down). We used the output vector at layer FC8 or FC9 as input to MLP. It is considered that output vector of FC8 is related to semantic attribute and that of FC9 is label of generic object. We called it Method1 of which takes the output vector of layer FC9 to the input of MLP, and Method2 which takes the output vector of FC8 as the input. In our proposed method, in addition to the mid-level feature, features of object parts extracted from CBN are employed. We merge these features, output vector of FC8 or FC9 of CNN and bottleneck feature of the object parts, and estimate the function using them as input to MLP. In training the specific object function, we collected positive images with the objects and negative images without the objects.

IV. Experiments

A. Dataset

In the experiment, we collected the images from ImageNet[17]. 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 5 classes (“containable”, “cuttable”, “movable”, ”writable”, ”sittable”).

We collected cup, kettle, paper cup, can, mug cup for “containable”. In the same way, knife, scissors, ax were collected for “cuttable” and bicycle, train, wagon and bus for

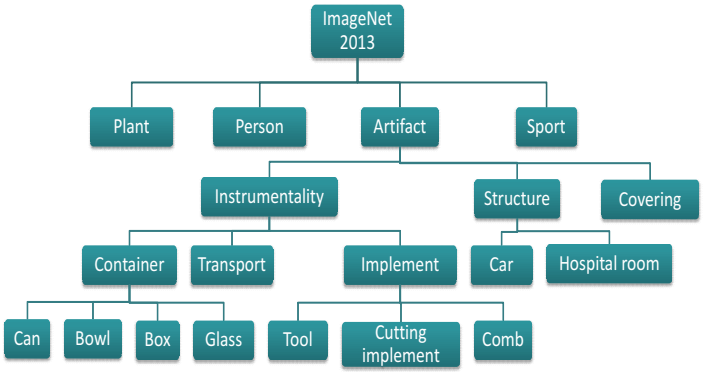

Fig. 6: Overview of WordNet

“movable” and pen, fountain pen for “writable”, and sofa, chair, stool for “sittable”(see Fig. 5).

This is because the above five functions can be expressed by appearance. Fig. 6 shows the overview of WordNet. The “containable” objects were collected from “container” node in WordNet, the “cuttable” objects from “implement” node, the “movable” objects from “transport” node, the “writable” objects form “writing implement” node, the “sittable” objects from “seat” 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 images was about 1000 per function class respectively.

B. Experimental condition

In this experiment, we used the OverFeat[18]. Overfeat is the CNN which was pre-trained using 1,281,167 images

TABLE I: Classification rates. (%)

	Method 1		Method 2		only
	previous	proposed	previous	proposed	parts feature
Containable	86.7	87.9	86.7	88.1	32.0
Cuttable	57.4	57.2	56.9	62.0	18.5
Movable	75.0	73.4	72.2	75.6	39.3
Writable	68.2	69.9	71.7	73.6	42.0
Sittable	60.4	44.8	37.1	58.5	6.6
Average	71.8	69.1	67.2	**73.5**	27.3

in the CLS-LOC dataset of ILSVRC2013. Number of layers and nodes with MLP in Fig.4 down are 2 and 500, 200 respectively. Number of components of DPM and of part filters is 4 respectively. In addition, we evaluate 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 regarded 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 shows the classification results by the proposed Method 1 and Method 2 as well as our previous method. In the figure, the results are also listed using only feature of object parts. By our method in Method 2, the average of classification result for estimation achieved the highest rate, 73.5%. On the other hand, the average of classification result by Method 1 is lower than the previous method. In case using only feature of object parts, the classification rate is lowest. Output vector of FC8 used in Method 2 is considered to be related to semantic attribute, and then combined with the feature of object parts, it becomes more effective for function estimation. Therefore, it may be concluded that only feature of object parts is not effective for function estimation, but combined with specific feature, it contributes to function estimation.

V. Conclusion and Future Work

Various elements independently or mutually express the object function. We believe that function is closely related to the appearance, especially not only whole the object but object parts, therefore we proposed the method that estimates object function focusing on CNN. Classification rate of our proposed method was improved by 5.3% compared with the previous Method 2. Also, compared with the highest classification rate by our previous Method 1, our proposed Method 2 is higher by 1.7%. Therefore, It can be concluded that feature of object parts contributes to function estimation when combined with object feature.

However, parts extracted by DPM aren't directly related to object function. In a future, we will estimate function by training CNN, and then identifying parts which are directly related to function by backtracing the CNN from output.

References

[1] Rosch, Eleanor, et al. "Basic objects in natural categories." Cognitive psychology 8.3, pp.382-439, 1976.

[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, pp.815-824, 2005.

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

[6] K. Veselý, M. Karafiát, and F. Grezl, "Convolutive bottleneck network features for LVCSR" in ASRU, pp.42-47, 2011.

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

[8] 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.

[9] Saxena, Ashutosh, Justin Driemeyer, and Andrew Y. Ng. "Robotic grasping of novel objects using vision." The International Journal of Robotics Research 27.2, pp.157-173, 2008.

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

[11] 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.

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

[13] 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.

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

[15] Pieropan, Alessandro, Carl Henrik Ek, and Hedvig Kjellström. "Recognizing Object Affordances in Terms of Spatio-Temporal Object-Object Relationships." Humanoid Robots(Humanoids), 2014 IEEE-RAS International Conference on Humanoid Robots on. IEEE, 2014.

[16] Kitano, Yosuke, Tetsuya Takiguchi, and Yasuo Ariki. "Estimation of object functions Convolutional Neural Network." Frontiers of Computer Vision (FCV), 2016 22nd Korea-Japan Joint Workshop on. IEEE, 2016.

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

[18] Sermanet, Pierre, et al. "Overfeat: Integrated recognition, localization and detection using convolutional networks." arXiv preprint arXiv:1312.6229, 2013.