Estimation of Object Functions Using Deformable Part Model

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, the most important thing is not the names but functions 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 DPM by dividing object appearance into parts. There are important parts and less important parts in the appearance for the functions. Therefore, we identify the important parts in the object appearance for the functions. Experimental results show that the important parts about specific functions can be extracted and object functions are related to the appearance.**

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, the most important thing is not to recognize the object name such as "pen", but to recognize the function allowing us to write. 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.

We show the example of basic level category and function level category of objects in Fig. 1. In this paper, recognizing

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

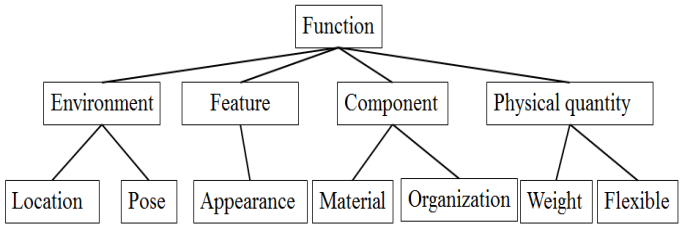

Fig. 2: Function-based ontology

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. Though the appearance in functional classes has a wide variety, there are common parts in the functional classes. For example, there is a "plate" as the common part in the function allowing us to sit on. Therefore, our goal is to identify the important parts forming the object functions. To this end, we create the functional models using

DPM(Deformable Part Model)[5], and estimate the object functions. In addition, we identify the most important part forming the specific function among the object parts separated by DPM. In the experiment, we test the unknown object images to evaluate the function estimation.

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], [10], [11], they define affordance as the relationship between robotics hand and objects, while according to [12], 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 [13], [14], they set up the task that makes the robot search for the object where humans can sit. In [15], humans might interact with the same object in different ways, with only some typical interactions corresponding to object affordance. [12], [16] 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, these researches limit the task or environment too much so that it seems better to perform the specific object recognition and to annotate the function label. In this work, we estimate the object functions without limiting the task or environment. If we estimate the object functions 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.

DPM is currently the state of the art for object detection. 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. This system achieves a two-fold improvement in average precision over the winning system in the 2006 PASCAL person detection challenge[17]. The system also outperforms the best results in the 2007 challenge in ten out of twenty object categories[18]. In this work, we use DPM to divide objects into parts and score each part.

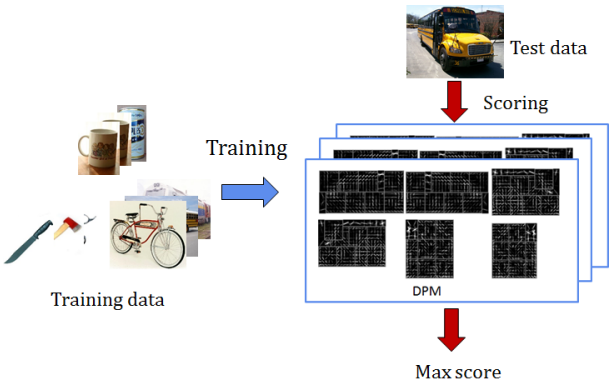

Fig. 3: Overview of function estimation using DPM.

III. Function estimation using DPM

An overview of our approach is shown in Fig. 3. We train models using [5], which learn latent parts given object bounding boxes. This model produces aspect mixture components and part configurations associated with detection. Our strategy involves training procedure of a single model per function class, using a various kind of object images with the function as positive and also images without the function as negative. $f^j(a)$ is the j-th function score and obtained as follows:

$$f^j(a) = F_0 \cdot \phi(H, p_0) + \sum_{i=1}^{n} \max_{p_i} (F_i \cdot \phi(H, p_i) - d_i \cdot \phi_d(dx_i, dy_i)) \tag{1}$$

where a is a test image, F_0 is a root filter, $p_i = (x_i, y_i, l_i)$ specifies the level of feature pyramid (l_i) and position(x_i, y_i) of the i-th filter, F_i is a filter for the i-th part. $\phi(H, p_i)$ denotes the vector obtained by concatenating feater vectors in the subwindow of H with top-left p_i in row-major order. Here, root filter approximately covers an entire object and part filters cover the parts of the object.

$$(dx_i, dy_i) = (x_i, y_i) - (2(x_0, y_0) + v_i) \tag{2}$$

gives the displacement of the i-th part relative to its anchor position. Here, v_i indicates a two-dimensional vector specifying an anchor position for part i relative to the root position. The following equation shows deformation feature.

$$\phi_d(dx, dy) = (dx, dy, dx^2, dy^2) \tag{3}$$

The score of a hypothesis is given by the scores of each filter at their respective locations minus a deformation cost that depends on the relative position of each part with respect to root, plus the bias.

As shown in Fig. 4, each function model scores the object images. Applying all function models to an object image a, the label $C(a)$ is predicted with the highest score which the corresponding function model reports. Here, when the score of an object image is lower than the threshold which was set up in training, we consider the object has no function.

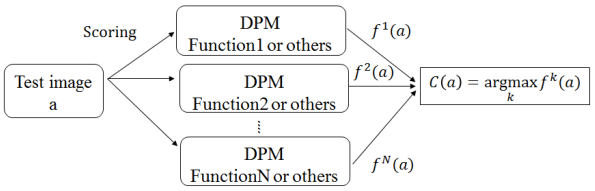

Fig. 4: Implementing the "one-vs-rest" approach for multi-class classification.

IV. Experiments

A. Dataset

Our goal is to estimate object functions and identify the functionally important part on the object. In this experiment, we collected the images from ImageNet[19]. It is an image database formed according to 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"). This is because a few functions can be expressed by appearance in the first place and it is assumed that "movable" can be expressed by the tires.

We prepared cup, kettle, can, mug for training, and pod for test as "containable", bicycle, train, wagon for training and bus for test as "movable" and knife, scissors, ax for training and punch for test as "cuttable" (see Fig. 5).

Fig. 5: Image examples in ImageNet

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" in WordNet, but we regard them as "movable" because they do not have the "containable" function. As the number of components of DPM, we tried 2, 4, 6 and 8. The number of training images and test images were about 1300 and 500 images per function class respectively.

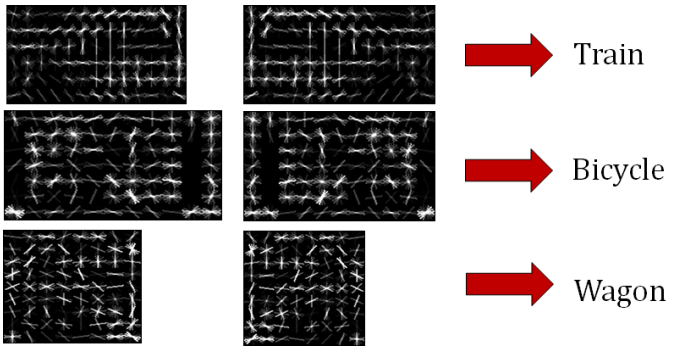

Fig. 6: Root model of "movable" trained using DPM

B. Experimental result

TABLE I to IV show the result of classification rate for the different number of the components. In spite of the number of components, "movable" function has the highest recognition rate. In addition, the highest classification rate of "containable" is 68.8% and the highest one of "cuttable" is 71.7%. Both of them are lower than the lowest one of "movable" function 73.6%. It can be seen from TABLE I to IV that "containable" is confused with "cuttable". Therefore, it can be said that "movable" function is well formed by the appearance.

Fig. 6 shows the model trained with 6 components for "movable" function, and each component corresponds to the object category. As the number of the components increases, the classification rate of "movable" is improved because the mixture component is the reliable and strong cue.

C. Identifying functionally important part

We identified the most functionally important part which forms the "movable" function. In Eq. 1, $F_i \cdot \phi(H, p_i)$ refers to the score of the i-th part. Therefore we regarded the part with the highest score as the most important one of the "movable" function. $score_{part}(i,k)$ gives the total score of the i-th part in k-th component.

$$score_{part}(i,k) = \sum_{l=1}^{L} F_{i,k} \cdot \phi(H, p_{i,l}) \tag{4}$$

where $F_{i,k}$ is a filter for the i-th part in k-th component and $\phi(H, p_{i,l})$ is $\phi(H, p_i)$ of the l-th example. As described in

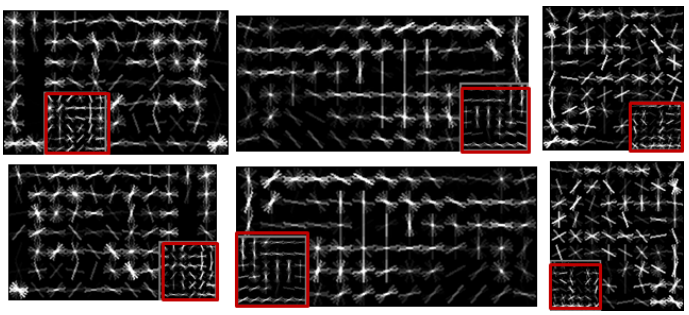

Fig. 7: Functionally important part of "movable"

TABLE I: Classification performance. The number of components is 2. (%)

	Containable	Cuttable	Movable
Containable	68.8	29.0	20.1
Cuttable	21.2	55.6	6.2
Movable	10.0	15.4	73.6

TABLE II: Classification performance. The number of components is 4. (%)

	Containable	Cuttable	Movable
Containable	65.1	19.3	17.1
Cuttable	27.0	69.7	8.3
Movable	7.8	11.0	74.6

TABLE III: Classification performance. The number of components is 6. (%)

	Containable	Cuttable	Movable
Containable	56.2	16.5	12.4
Cuttable	37.4	71.7	10.5
Movable	6.8	11.8	77.1

TABLE IV: Classification performance. The number of components is 8. (%)

	Containable	Cuttable	Movable
Containable	49.8	8.3	9.1
Cuttable	30.7	65.7	6.6
Movable	19.5	26.6	84.3

Section 3, we trained the three function models using DPM and scored the object images with "movable" function. In scoring an object image, the model chose the component with the highest score.

Fig. 7 shows the part with the highest score in each component. As shown in Fig. 7, the tire is found to be the most important functional part for "movable".

V. Conclusion and Future Work

Various elements independently or mutually express the function. We believe that function is closely related to the appearance, so we proposed the method that could estimate the object function using DPM and identified the functionally important part. Function estimation of "movable" class had 84.3% accuracy in the experiments. Our experiments have shown that "movable" function could be formed by appearance owing to the tire.

However, the function expressed by appearance is limited, because there are object properties not observed visually from a single image, such as flexibility, weight and use.

In the future, the method of function estimation will be extended in two ways. Firstly, CNN (convolutional neural networks) will be employed to find the attributes effective to the functional classification, because it shows state-of-the-art performances in many benchmark tests, such as object category recognition, handwritten character recognition. Secondly, various attributes like physical information [20] such as weight and length will be employed.

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] Bohg, Jeannette, and Danica Kragic. "Grasping familiar objects using shape context." Advanced Robotics, 2009. ICAR 2009. International Conference on. IEEE, 2009.

[9] Madry, Marianna, Dan Song, and Danica Kragic. "From object categories to grasp transfer using probabilistic reasoning." Robotics and Automation (ICRA), 2012 IEEE International Conference on. IEEE, 2012.

[10] 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.

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

[12] 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.

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

[14] 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.

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

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

[17] Everingham, Mark, et al. "The pascal visual object classes challenge 2006 (voc 2006) results." (2006).

[18] Everingham, M., et al. "The PASCAL Visual Object Classes Challenge 2007 (VOC 2007) Results (2007)." URL http://www. pascal-network. org/challenges/VOC/voc2007/workshop/index. html. 2008.

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

[20] Zhao, Yibiao, and Song-Chun Zhu. "Scene parsing by integrating function, geometry and appearance models." Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013.