Estimation of Object Functions Using Visual Attention

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
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 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 our previous work, we showed that object functions were closely related to the appearance. In this paper, we propose a new method to estimate object functions by focusing object parts. In this work, we estimate object function using visual attention model, then visualize regions of image, which contribute to predict object functions. Experimental results show that the classification rate of five functions is improved by 0.5% compared with the previous method.**

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 manages 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.

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

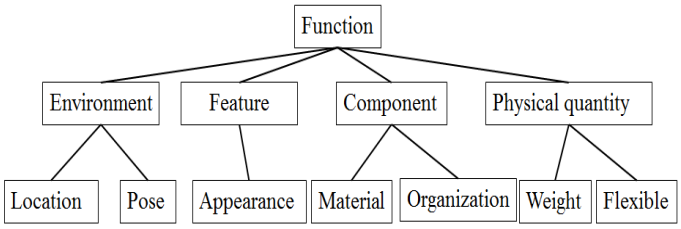

Fig. 2: Function-based ontology

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 are 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.

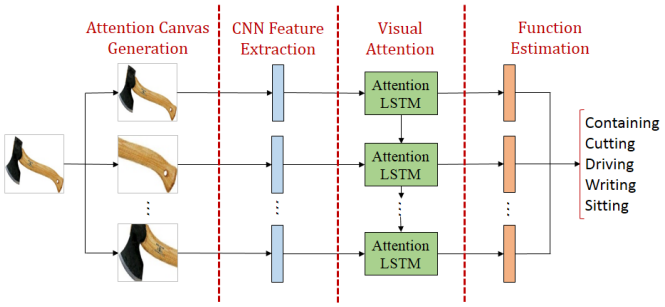

Fig. 3: Overview of proposed method

In our previous work[5], we showed that object functions are closely related to object parts. However, we couldn't find important object parts which are directly related to object function. In this work, we estimate object function using visual attention model, then visualize regions of image, which contribute to predict object functions. In addition, to find multiple attentive regions, we utilize attention canvases[6].

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 our previous work[5], 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. In addition to the mid-level feature of CNN, we used feature of object parts extracted from Deformable Parts Model(DPM)[16] and Convolutive Bottleneck Network(CBN)[17].

However, we couldn't find important object parts which were directly related to object function. Therefore, in this work, we visualize object parts which are related to object function by using visual attention model. Visual attention model is used for caption generation[18], action recognition[19], et al. Visual attention model shows where the model is focusing its attention. In [6], they generated attention canvases by cropping images according to various size of windows and stride, then applied visual attention model. In this work, we visualized regions of object image, which were related to object function, by generating attention canvases.

III. Function Estimation using Visual Attention

An overview of the propoesd method is shown in Fig. 3. It is composed of four steps shown below from A to D.

A. Attention Canvas Generation

Firstly, attention canvas is generated. As attention canvas, images of dataset are cropped according to various defined size of windows and stride. Then, these attention canvases are normalized to the uniform size.

B. CNN Feature Extraction

Secondly, convolutional neural network feature is extracted. Convolutional neural network is trained for generic object recognition. CNN feature is extracted from attention canvas, at the last convolutional layer.

C. Visual Attention Model

Since we want to find different attentive regions in an image, the attention canvas are fed into LSTM[20]. LSTM with attention model is shown in Fig. 4.

The feature map extracted at time t is represented as follow: $\boldsymbol{X}_t = [\boldsymbol{X}_{t,1}, \cdots, \boldsymbol{X}_{t,K}, \boldsymbol{X}_{t,K+1}, \cdots, \boldsymbol{X}_{t,K^2}]$

With feature map $\boldsymbol{X}_t$ and hidden state of previous LSTM unit $\boldsymbol{h}_{t-1}$, the new attention map is defined as follows:

$$l_{t,i} = \frac{\exp(W_{h,i}^{\mathrm{T}} \boldsymbol{h}_{t-1} + W_{x,i}^{\mathrm{T}} \boldsymbol{X}_t)}{\sum_{j=1}^{K^2} \exp(W_{h,j}^{\mathrm{T}} \boldsymbol{h}_{t-1} + W_{x,j}^{\mathrm{T}} \boldsymbol{X}_t)}, \forall i = 1, \ldots, K^2,$$

where $W_{h,i}$ refers to weights of the connections from previous hidden state $\boldsymbol{h}_{t-1}$ to the i-th location of the spatial attention map. Similarly, $W_{x,i}$ denotes the weights from feature map $\boldsymbol{X}_t$ to the i-th location of the map. Then, the attentive feature $\mathbf{x_t}$ is calculated by the weighted summation over the feature map $\mathbf{X_t}$ based on the predicted attention map l_t:

$$\boldsymbol{x}_t = \sum_{i=1}^{K^2} l_{t,i} \boldsymbol{X}_{t,i}$$

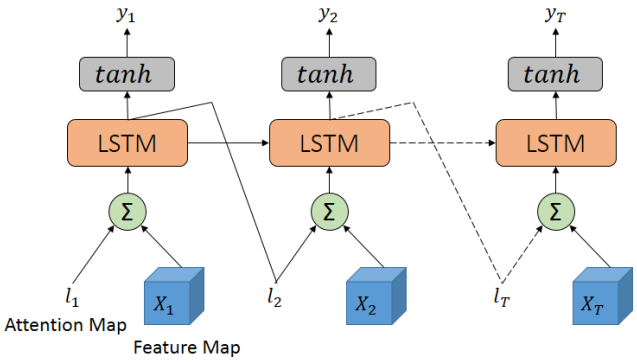

Fig. 4: Recurrent model with attention

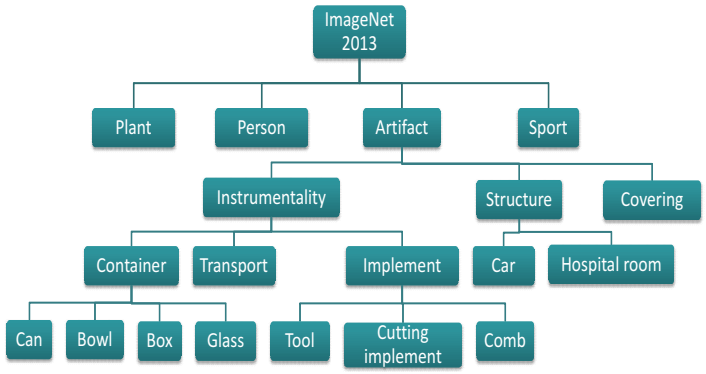

Fig. 5: Overview of WordNet

The loss function is defined as follows:

$$\boldsymbol{L} = -\sum_{t=1}^{T}\sum_{i=1}^{C} y_{t,i}\log\hat{y}_{t,i} + \lambda\sum_{j=1}^{K^2}(1-\sum_{t=1}^{T} l_{t,j})^2$$

where $y_{t,i}$ is an output label vector, T is the total number of time steps, and λ is an attention penalty coefficient.

D. Function Estimation

Finally, object function is estimated. The hidden state h_t of LSTM followed by a $tanh$ activation function is used as the features for classification. The final object function classification result is the average of the classification results across all time steps.

IV. Experiments

A. Dataset & Experimental condition

In the experiment, we collected the images from ImageNet[21]. 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 we collected the images from ImageNet is that we can associate functions with synsets.

The task of function estimation is carried out for 5 classes ("containing", "cutting", "driving", "sitting", "writing"). We collected cup, glass, punch bowl, bottle, vessel, tea kettle for "containing". In the same way, knife, scissors, ax, wire cutter were collected for "cutting" and bicycle, motor scooter, car and bus for "driving" and pencil, crayon, marker, quill pen, fountain pen for "writing", and sofa, chair, bench, ottoman, stool for "sitting".

This is because the above five functions can be expressed by appearance. Fig. 5 shows the overview of WordNet. The "containing" objects were collected from "container" node in WordNet, the "cutting" objects from "implement" node, the "driving" objects from "transport" node, the "writing" objects from "writing implement" node, the "sitting" objects from "seat" node in WordNet. The number of images was about 250 per function class respectively.

In this experiment, we used OverFeat[22] which was trained using 1,281,167 images in the CLS-LOC dataset of ILSVRC2013. In addition, we evaluated our proposed model on semi-closed condition, not using cross-validation because of large training time.

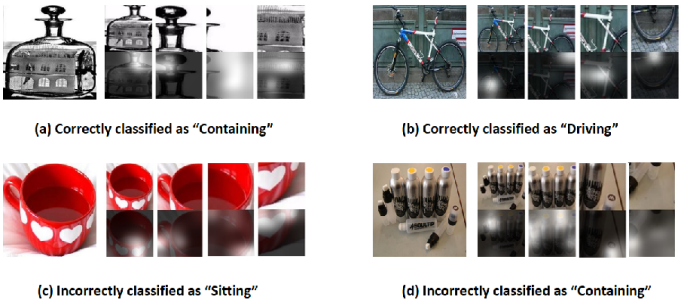

Fig. 6: Visual attention example.($\lambda = 1$)

B. Experimental result

TABLE I shows the classification results by the proposed method and our previous method, carried out on same condition. Here, λ indicates an attention penalty coefficient. We executeed previous method, and compared it with the proposed method on the same condition. By the proposed method($\lambda = 0$), the averaged estimation achieved the highest rate, 85.2%. On the other hand, when $\lambda = 1, 10$, the proposed method is lower than the previous method. Fig.6 shows some attention canvases of dataset, and its attention visualization. In Fig.6(c), cups which have "containing" function are incorrectly classified as "sitting" function. In this example, visual attention maps show strong attention to the texture of surface. When λ becomes large, the loss function begins to depend on attention penalty item. Therefore, when the model attends to regions which don't relate to function, the larger λ becomes, the lower classification rate is. On the other hand, in Fig.6(a), the model attends to spout which relates to "containing" function. In Fig.6(b), the model attends to wheel which relates to "driving" function.

TABLE I: Classification rates. (%)

	Proposed method			Previous method
	$\lambda = 0$	$\lambda = 1$	$\lambda = 10$	
Containing	88.2	83.7	83.9	89.7
Cutting	73.9	72.4	65.6	73.5
Driving	96.6	94.9	95.6	95.0
Sitting	82.8	82.6	82.5	82.1
Writing	84.4	85.2	80.8	83.2
Average	**85.2**	83.8	81.7	84.7

V. Conclusion and Future Work

Various elements present the object function independently or mutually. We think that function is closely related to the appearance of, especially not only whole the object but object parts. From this viewpoint, we proposed the function estimation method that attends to the object parts. Classification rate of our proposed method(λ=0) was improved by 0.5% compared with the previous method. In addition, we attend to the region which related to object function. However, when attention penalty coefficient λ is large, out model strongly attends to regions which don't relate to object function. In a future, by increasing the number of attention canvas, we try to attend more small region which is related to object function.

Acknowledgment

A part of this study is subsidized by JSPS Grant-in-Aid for Scientific Research and Research granted JP 17K00236.

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] Azuma, Ryunosuke, Tetsuya Takiguchi, and Yasuo Ariki. "Estimation of object functions Focusing on Feature of Object Parts." International Workshop on Frontiers of Computer Vision (IW-FCV), 2017 23rd International Workshop on. IEEE, 2017.

[6] Zhao, Bo, et al. "Diversified Visual Attention Networks for Fine-Grained Object Classification." IEEE Trans. Multimedia, vol 19, no.6, pp. 1245-1256, Jun, 2017.

[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] 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.

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

[18] K. Xu, J. Ba, R. Kiros, et al. "Show, attend and tell: Neural image caption generation with visual attention." in ICML, 2015, pp. 2048-2057.

[19] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention.", in NIPS Time Series Workshop, Dec. 2015.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory."" Neural Comput., vol. 9, no. 8, pp. 1735-1780,Nov. 1997.

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

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