Automatic Construction of Image Dataset from Web using Ontology and Similarity of Images

Takahiro Yoshimura
Department of Computer Science
Chubu University
Kasugai, 487-8501 Japan
Mail:takatecep@gmail.com

Yuji Iwahori
Dept. of Computer Science
Chubu University
Kasugai, 487-8501 Japan
Mail:iwahori@isc.chubu.ac.jp

Boonserm Kijsirikul
Dept. of Computer Engineering
Chulalongkorn University
Bangkok, 10330 Thiland
Mail:Boonserm.K@chula.ac.th

Shinji Fukui
Faculty of Education
Aichi University of Education
Kariya, 448-8542 Japan
Mail:sfukui@auecc.aichi-edu.ac.jp

Yoshitsugu Hayashi
Institute of Sci. and Tech. Research
Chubu University
Kasugai, 487-8501 Japan
Mail:y-hayashi@isc.chubu.ac.jp

Witsarut Achariyaviriya
Dept. of Constructional Engineering
Chubu University
Kasugai, 487-8501 Japan
Mail:witsarut.ac@gmail.com

***Abstract*—Dataset is important for the general object recognition and this paper proposes an automatic construction of image dataset from Web images. Web image mining approach is introduced to construct image dataset where noisy images exist among web images although the collecting cost is low. The proposed approach uses meta information added to the image and tries to collect more images using Ontology and similarity of features between the training image and collected image to remove the noisy images. The paper aims the automatic collection of image dataset by removing the noisy images with high accuracy. The results suggest Precision 94.0%, Recall 84.1% and F-measure 88.6% for the collected image dataset in the experiments. It is shown that the proposed approach collects various aspects of image data to be applied to the object recognition.**

***Index Terms*—Ontology, Noisy Image Removal, Image Dataset, Similarity**

I. Introduction

General object recognition is one of the representative task which recognizes object using the real images in the image recognition. In 2000s, feature vectors such as SIFT[1] or HOG[2] developed by researchers and combination of those feature vectors with classifier such as SVM[3] have been proposed. Competition of image recognition named "The ImageNet Large Scale Visual Recognition Challenge[4](以下,ILSVRC)" have been held by ImageNet[5] since 2010 and recognition errors were over 25% in 2010 and 2011. CNN (Convolutional Neural Network) has been proposed since 2012 and error ratio was reduced to 15.3%. New CNN models have been proposed at ILSVRC and CNN approach (i.e., Deep Learning) has been main approach for image recognition since 2012. The recognition ratio of CNN has exceeded that of human recently. Improvement of recognition accuracy is based on the evolution of algorithm and processing performance and dataset to be trained for CNN. How to collect dataset is one of important issues these days.

When number of training samples in the dataset exceeds some threshold by the previous approaches before Deep Learning, recognition ratio has some limitations, while Deep learning approaches can improve the recognition ratio logarithmically proportional to the number of training samples.

Object detection, segmentation and pose estimation are representative tasks using CNN and these networks and object recognition network which got the high score in ILSVRC is used to extract features as the base network. Ref.[6] shows the importance of increasing the recognition accuracy according to increasing the number of training samples using big dataset for each task of object detection, segmentation and pose estimation. About the construction of dataset, there is an approach by hand made and searching Web image using Web image mining. Image dataset by hand made reflects the intention of constructor while there are many images taken by many persons in the Web. This makes it possible to obtain many real images with small cost. Web image mining takes lower cost to collect many images than construction by hand made but usually there are some problems that some noisy images not appropriate to the purpose are included in the collected images from Web. To solve this kind of problem, Ref.[7] and Ref.[8] removes the noisy images using the center of gravity

obtained from collected images in the process of Web image mining. This approach removes noisy images when the portion of noisy images are small among the collected images but there are some problems to remove noisy images when the portion of noisy images are large since center of gravity is determined from the features obtained from noisy images. Taking center of gravity is sometimes problem to collect appropriate images by the provided keyword. This paper proposes the construction of highly accurate dataset using removal of noisy images by the similarity of CNN features obtained from teaching images and collected images respectively.

II. Construction of Image Dataset

Proposed approach constructs highly accurate image dataset by removing noisy images for the collected images using ontology and similarity between images. Proposed approach constructs image dataset as follows.

Step 1 Collection of Images from Web

Step 2 Removal of Duplicate Images

Step 3 Removal of Noisy Images of collected images

A. Collecting Images from Web

Step 1 consists of collection of images from Web based on the search query using the label of dataset to be constructed. Proposed approach collects images via Web API using Flickr as image share site. Searching image is possible by the search keyword by Flickr. Images uploaded in Flickr has tags which means the contents of images and were are added when contributor uploads image. Proposed approach searches images with tags attached to those using search query and then collects images with specified tags.

Search by Sub-concept

Image search from Flickr site uses tags attached by users. When the label of image dataset is ”Dog”, some tags attached to the image cannot be hit even if the objective label is included in the image. Example is shown in Fig.1. However if the tag attached to the image is the word Chihuahua in the subordinate concept, it is a subset of the word Dog as the search query that represents the target label. Therefore, using the subordinate concept of the search query obtained from the ontology that is a knowledge system, images are collected again to increase the collected images.

Proposed method uses DBpedia [9] which is an ontology constructed from Wikipedia as its information source. Query language SPARQL is used from DBpedia to obtain the sub-concepts of search queries. Acquired subordinate concept words is used as search query and images are collected again.

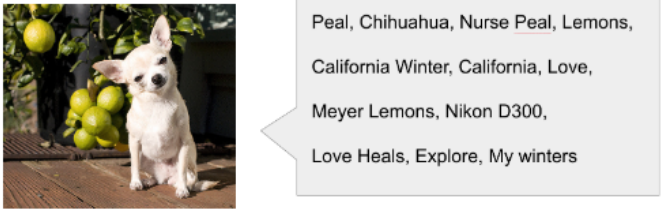


Fig. 1. Tags Attached to Image

Search in Multiple Languages

On Flickr, there are images uploaded by users from various countries and attached tags are written by multiple languages. Images are collected by multiple languages as search queries to correspond to the tags attached with multiple languages.

B. Removal of Duplicate Images

Images collected in Step 1 include the words of search query and those of search query subordinate concept or both of them as tags. Duplicate images are collected by this cases. Removal processing of duplicate images are performed to avoid duplication of images in a dataset. Proposed method uses a hash function to remove duplicate images. If the input values of the hash function are the same, the deterministic feature is used so that the same output value is obtained. When hash values take the same, it is judged that duplicate images were obtained and those should be removed.

There are some cases that images on the web have the same contents but sizes are different and those have been uploaded. These images represent the same content from a human vision, but are different data from a computer vision. Common hash functions such as MD5 and SHA-1 have a problem that duplication is not detected. Therefore, the proposed approach uses pHash [10] as one of Perceptual Hash which is a hash function for media data such as images and sounds.

C. Removal of Noisy Images from Collected Images

The images collected from the Web contain images that cause noise that is not suitable for search queries. These noisy images can not be applied to the general object recognition. So removal processing of noisy images is performed on the collected images.

The flow of removal of noisy images processing is as follows.

Step 1 Extract feature from teaching image

Step 2 Feature extraction from collected images

Step 3 Similarity calculation between features

Step 4 Judgment of noisy image by threshold

Previous papers [7] [8] proposed methods that remove the noisy images using the center of gravity of the features obtained from the collected images. These previous methods have a property that the removal processing works well when

the ratio of the noisy images included in the collected images is small. When many noisy images are included, the center of gravity is also based on the features obtained from the collected noisy images. In this case, removal of noisy images is not performed properly and it depends on the properties of the collected image. Here, this approach proposes a method that removal of noisy image does not depend on the properties of the collected images.

Any number of teaching images are prepared corresponding to the labels of the dataset to be constructed, The more teaching images are, the better it becomes but the processing time takes longer. Feature extraction is performed from the training images corresponding to the label of the dataset to be constructed. CNN is used as a feature extractor in the proposed method. Inception-v3 []DBLP:journals/corr/SzegedyVISW15 is used as a CNN for feature extraction.

Inception-v3 is a CNN model of a development of the GoogleNet architecture which is the winning model of the 2014 ILSVRC. Inception module is a feature of Inception-v3 and it is a small network composed of multiple convolutional layers and pooling layers. It has a structure of 1×1, 3×3, 5×5 convolution and pooling are stacked in parallel on the same layer. This structure is a Inception module and a single CNN is constructed by stacking the Inception modules. Just before the 3×3, 5×5 convolutions, and just after pooling, a 1×1 convolution layer is inserted, reducing the number of input channels in the subsequent convolution layers. As a result, number of weights and number of operations can be reduced.

Proposed method uses the output of the Pooling layer (2048 dimensions) obtained from the vicinity of the Inception-v3 final layer as features extraction. Similarly, features are extracted from each collected image and cosine similarities between features are obtained. If the similarity is greater than or equal to the threshold value, it is judged a non-noisy image.

III. Experiment

Here, dataset is constructed by the proposed method by computer experiments. Evaluation and recognition experiments are conducted and evaluated to confirm the effectiveness of the proposed method.

A. Evaluation Experiment

Using datasets obtained by the proposed method, experiment for evaluation was performed to confirm whether an appropriate label was attached or not. The experimental conditions are shown below.

Experimental Conditions

The same class as the general object recognition dataset CIFAR-10 [11] is used as the dataset to be constructed. CIFAR-10 is a subset of 80 million tiny image and this dataset is used for general object recognition where about 60,000 images are extracted and labeled. Class labels consist of 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Images used were collected from the class labels with queries using the Flickr API. 100,000 images were collected for each class and 100 images were randomly selected out of 100,000 images. A total of 1,000 images per each class were used in the experiment. Class label for the query at the time of collection was used as the label of the collected image. In the experiment, 10 types of classes were evaluated using the application for evaluation shown in Figure 3. 9 evaluators who are not related to this research evaluated whether the collected images were appropriate for each label or not. Images that were evaluated as appropriate by more than half of the evaluators are taken as correct images. Threshold for removing noisy image was set to be 0.65. In addition, 20 images collected manually for each class were used as the teaching image used for similarity determination. The evaluation of the experiment was performed using the following equations of Precision, Recall and F-measure, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

Precision for each class at the time when image dataset were collected from the Web is shown in Table I. Example images at the time of collection are shown in Figure 4.

Figure 4 shows that images where the target object are captured from various angles, images which is quite unrelated to the target object, minority images such as illustrations of object or dolls are included. Consideration suggests that these images include the possibility that features of the object may be incorrectly learned and these images should not be included in the image dataset for the training.

TABLE I
Evaluation of Accuracy for Each Label at Time of Collection from Web

Search Keyword	Precision[%]
airplane	81.0
automobile	77.0
bird	85.0
cat	73.0
deer	53.0
dog	73.0
frog	60.0
horse	67.0
ship	68.0
truck	38.0
AVG	67.5

B. **Experimental Results**

Experimental evaluation results after removing noisy images is shown in Table II. Examples of Non-noisy image and noisy image are shown in Fig.5 to Fig.14.

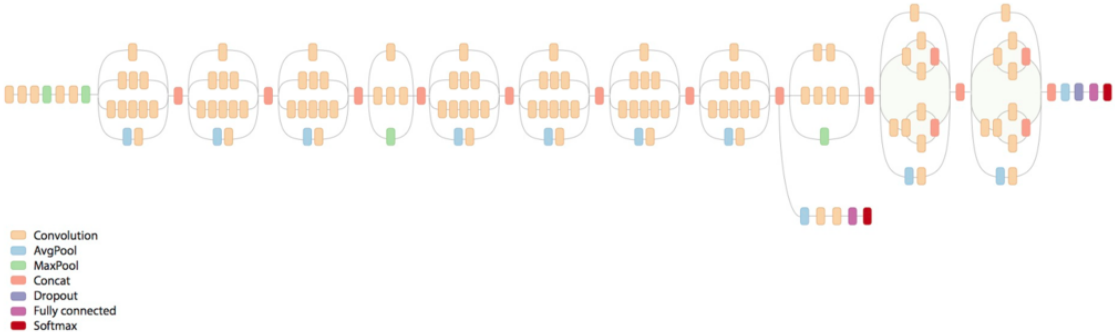


Fig. 2. Inception-v3

Fig. 3. Application for Evaluation

Fig. 4. Example Images at the time of Correction

Table II shows that Precision of 90.0 % or more is confirmed in all 9 labels except horse and it is shown that high accuracy could be obtained.

In all nine labels except horse It can be confirmed that it has a Precision of 90.0 % or more, It can be seen that it has high accuracy. While the Precision at each label is high, Focusing on recall of dog, the recall is 67.6 % and this is lower than other labels. Many non-noisy images were included among the group judged as noisy image. This may be based on the fact that feature extractor of Inception-v3 is pre-trained with ImageNet and dogs are subdivided with further labels. Teaching images used in this experiment could not cover these multiple labels. Recall is about 10 % higher for the label bird but the same thing is considered.

In the image examples of non-noisy images, it is confirmed that plane, automobile, bird, cat, frog, ship and truck have a relatively clear as a target. While it is confirmed that deer, dog and horse show that target object is captured but the background has more region or different objects (people, bicycles, playground equipment) were included. It is confirmed that noisy images of bird, cat, deer, horse include minority images such as origami, dolls and character images of target object and target images are properly judged as noisy images.

C. Experiment of Constructing Tuktuk Dataset

An experiment of constructing the Tuktuk dataset was done to confirm whether the proposed method can construct a dataset of objects which exist in Thailand. Tuktuk images were gathered by Flickr API with Tuktuk as a query word. As a result, 16915 images were obtained. 100 images were randomly chosen from the images to use the experiment of removing noisy images. 20 images in 16915 images were selected manually as the correct images. The threshold for removing noisy images is set to 0.8. Many noisy images exist in the images obtained from Flickr, and there are a few varieties of images of Tuktuk. The threshold is set higher to obtain the Tuktuk images with high accuracy. Other conditions were the same as the above experiment.

The examples of the experimental results are shown in Fig. 15. Precision, Recall, and F-measure are 100%, 39.6% and 56.7%, respectively. The result shows that the proposed method can construct the Tuktuk dataset from the Flickr web site with high accuracy because the Precision is 100% even though Recall is low.

D. Recognition Experiment

Validity of the image dataset constructed by the proposed method was evaluated. Dataset before removing noisy images and dataset after removing noisy images, the recognition

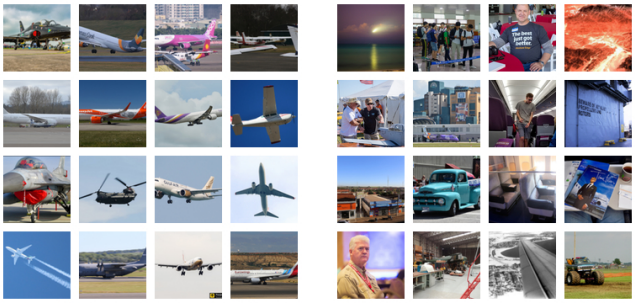
rate for the object recognition was evaluated under the same conditions with a manually generated image dataset. CIFAR-10 was used as dataset for comparison.

100,000 images were collected for each class using the proposed method. Learning images

5,000 images were randomly sampled from the images before removing noisy image and used for training images before removing noisy images. Also 5,000 images were randomly sampled from non-noisy images were used for training images after removing noisy images. The threshold and teaching images used for removing noisy images were the same as those used in the previous section. Figure III shows the number of images judged as non-noisy images. 1,000 test images prepared from CIFAR-10 were used for each class. Test images were commonly used to all datasets.

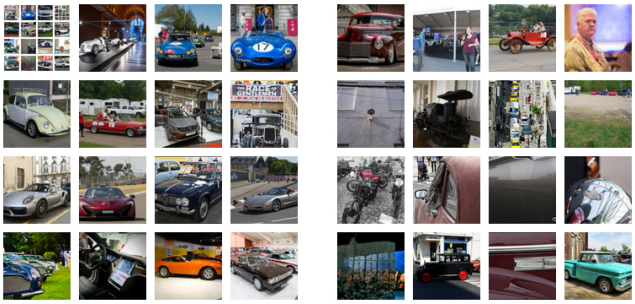
TABLE II
RESULTS OF EVALUATION IN EXPERIMENT

Search Keyword	Precision	Recall	F-measure
airplane	92.9	98.7	95.7
automobile	92.4	81.3	86.5
bird	97.0	78.3	86.7
cat	98.4	84.5	91.0
deer	91.7	84.6	88.0
dog	92.3	67.6	78.0
frog	96.2	84.7	90.0
horse	89.7	93.8	91.7
ship	92.7	76.1	83.6
truck	97.1	91.7	94.3
AVG	94.0	84.1	88.6



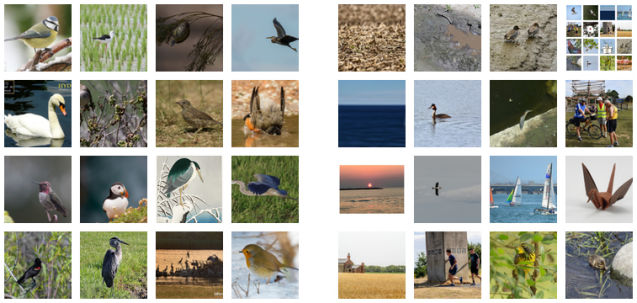
Non-Noisy Image Noisy Image

Fig. 5. airplane



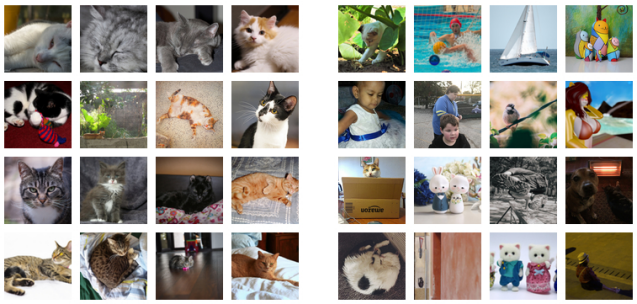
Non-Noisy Image Noisy Image

Fig. 6. automobile



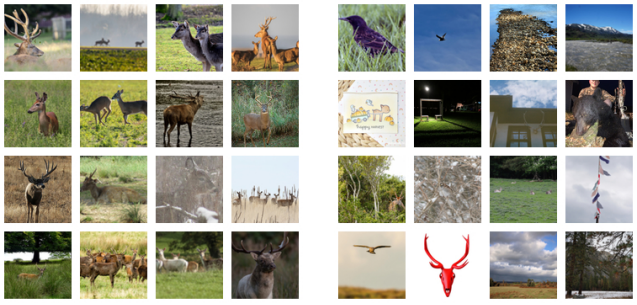
Non-Noisy Image Noisy Image

Fig. 7. bird



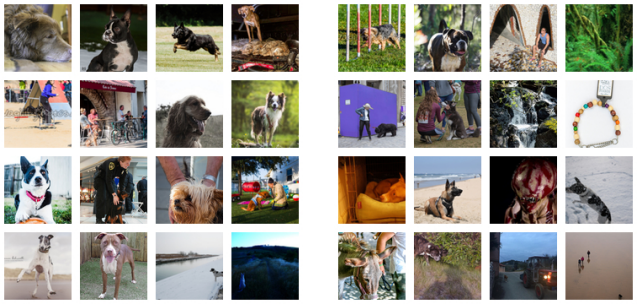
Non-Noisy Image Noisy Image

Fig. 8. cat



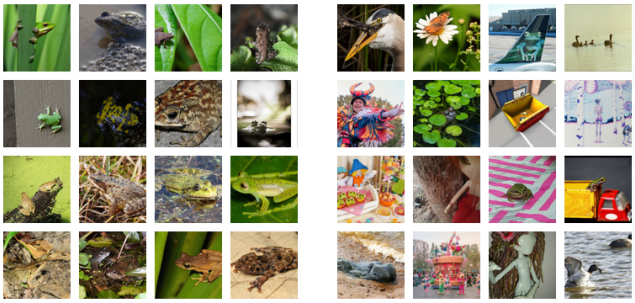
Non-Noisy Image Noisy Image

Fig. 9. deer



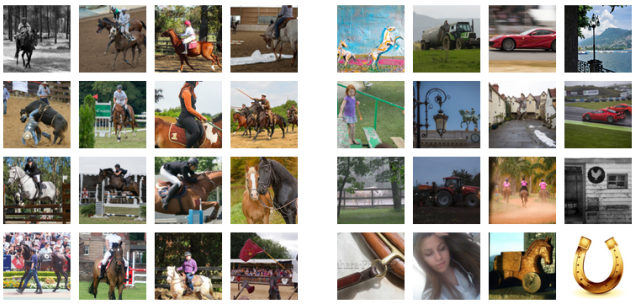
Non-Noisy Image Noisy Image

Fig. 10. dog



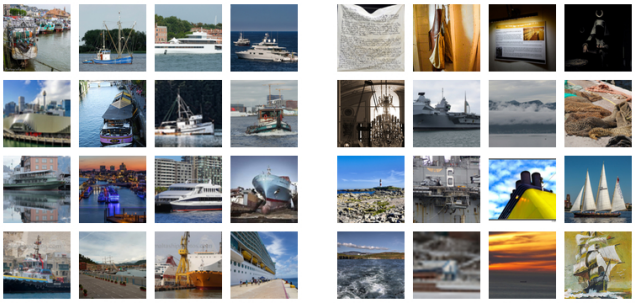
Non-Noisy Image Noisy Image

Fig. 11. frog



Non-Noisy Image Noisy Image

Fig. 12. horse



Non-Noisy Image Noisy Image

Fig. 13. ship

Non-Noisy Image Noisy Image

Fig. 14. truck

Non-Noisy Image Noisy Image

Fig. 15. Tuktuk

The network shown in Fig.16 was constructed and each of three datasets was trained and the recognition rate was obtained. Categorical Cross-Entropy was used for the loss function and Adam was used for optimization. Learning was done under the condition that initial value of learning rate was 1e-4, a batch size was 32 and a learning epoch was 20. This experiment was executed 5 times and average recognition rate was compared.

TABLE III
NUMBER OF NON-NOISY IMAGES JUDGED FROM 100,000 COLLECTED IMAGES

Label	Number of Images
airplane	81930
automobile	87345
bird	66767
cat	57516
deer	55312
dog	53007
frog	49988
horse	59929
ship	55354
truck	35454
AVG	60260

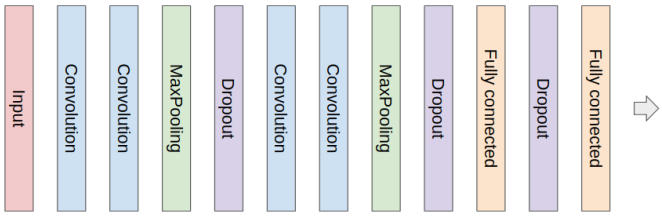


Fig. 16. Network used in Experiment

E. Experimental Results

Results of recognition performance are shown in Table IV. Table IV shows that the average accuracy has increased by around 25 % before and after removing noisy images. In the evaluation experiment in the previous section, images were collected with average of 94.0 % Precision. In the recognition experiment of this experiment, the recognition accuracy was 65.9 %, which was about 8 % lower than manually constructed dataset. This paper also suggests that Tuktuk images are

collected automatically to use as dataset for recognition. CNN approach becomes very popular and next stage of the research is to recognize many target objects including Thai oriented object. We expect CNN can estimate QoL (Quality of Life) with many collect images for the learning.

Test images of CIFAR-10 has the feature that only the target object is large and clear. While learning dataset constructed by the proposed method includes the feature that other objects are included except the target object or that the target object is small. It is considered that such difference between the learning image and the test image causes these results. However, human can recognize objects correctly even in such cases. Although the accuracy is lower than that of manually constructed image dataset, it was confirmed that various images were collected automatically excluding subjectivity.

TABLE IV
RESULT OF RECOGNITION EXPERIMENT

Dataset	Accuracy[%]
CIFAR-10	74.1
Proposed Approach (before removing noisy images)	40.6
Proposed Approach (after removing noisy images)	65.9

IV. CONCLUSION

This paper proposed a method of dataset construction using ontology and similarity between images. The range of collected images were expanded by using ontology conceptual relationships. Removing the noisy image using the similarity obtained from the CNN features of the teaching image and the collected image could construct non-noise image dataset, High accurate image dataset was constructed without depending on the ratio of noisy images in the collected images. Result of the evaluation experiment gave 94.0 % Precision, 84.1 % Recall and 88.6 % F-measure, respectively for collected dataset of non-noisy images.

Although the accuracy is lower than that of manually constructed image dataset, it was confirmed that collected image dataset has various images were collected.

Experiment on a dataset for general object recognition was done but developing an ontology specific to a domain and fine-tuning the network as a feature extractor makes it possible to construct dataset for the various domains of target.

Automatic judgment with appropriate thresholds and efficient selection of teacher images are remained including improving accuracy and judgment for removing noisy image.

ACKNOWLEDGMENT

This research is supported by SATREPS Project of JST and JICA: "Smart Transport Strategy for Thailand 4.0 Realizing better quality of life and low-carbon society", by Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (C) (17K00252) and by Chubu University GrantGrant. The authors would like to thank Mr. Naoki Watanabe for his experimental help.

REFERENCES

[1] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110, November 2004.

[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pp. 886–893, Washington, DC, USA, 2005. IEEE Computer Society.

[3] Marti A. Hearst. Support vector machines. *IEEE Intelligent Systems*, Vol. 13, No. 4, pp. 18–28, July 1998.

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, Vol. 115, No. 3, pp. 211–252, 2015.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[6] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *CoRR*, Vol. abs/1707.02968, , 2017.

[7] Seiki Otani, Ryosuke Yamanishi, and Yuji Iwahori. Generation of web image database based onhybrid noise removal method of visual and semantic features. *Transactions of the Japanese Society for Artificial Intelligence (in Japanese)*, Vol. 32, No. 1, pp. WII-N_1-10, 2017.

[8] Takahiro Yoshimura, Yuki Mizoguchi, Yuji Iwahori, and Ryosuke Yamanishi. Automatic construction of large scale image data set from web using ontology and deep learning model. *PATTERNS 2018*, pp. 21–22, 2018.

[9] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, Vol. 6, No. 2, pp. 167–195, 2015.

[10] Christoph Zauner and Martin Steinebach. Implementation and benchmarking of perceptual image hash functions, first edition. 2010.

[11] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).