International Journal of Networked and Distributed Computing, Vol. 3, No. 4 (November 2015), 234-242

Hybrid Approach of Ontology and Image Clustering for Automatic Generation of Hierarchic Image Database

Ryosuke Yamanishi [1], Ryoya Fujimoto [2], Yuji Iwahori [2] and Robert J. Woodham [3]

[1] College of Information Science and Engineering, Ritsumeikan University, 1-1-1 Nojihigashi,Kusatsu, Shiga 525-8577, Japan

E-mail: ryama@media.ritsumei.ac.jp

[2] Dept. of Computer Science, Chubu University, 1200 Matsumoto-cho, Kasugai, Aichi 487-8501, Japan

E-mail: iwahori@cs.chubu.ac.jp

[3] Dept. of Computer Science, University of British Columbia, 2075 Wesbrook Mall, Vancouver, B.C. Canada V6T 1Z4

E-mail: woodham@cs.ubc.ca

Abstract

This paper proposes a hybrid approach of ontology and image clustering to automatically generate hierarchic image database. In the field of computer vision, "generic object recognition" is one of the most important topics. Generic object recognition needs three types of research: feature extraction, pattern recognition, and database preparation; this paper targets at database preparation. The proposed approach considers both object semantic and visual features in images. In the proposed approach, the semantic is covered by ontology framework, and the visual similarity is covered by image clustering based on Gaussian Mixture Model. The image database generated by the proposed approach covered over 4,800 concepts (where 152 concepts have more than 100 images) and its structure was hierarchic. Through the subjective evaluation experiment, whether images in the database were correctly mapped or not was examined. The results of the experiment showed over 84% precision in average. It was suggested that the generated image database was sufficiently practicable as learning database for generic object recognition.

Keywords: Image database, Web intelligence, Image clustering, Cross media hybrid approach

1. Introduction

In the field of Computer Vision, "generic object recognition" is one of the most important and difficult topics. Generic object recognition tries to recognize object in a given image without any constraints. Some studies have tried to this problem with varied approaches [1,2,3]. Generic object recognition needs three types of elemental researches: feature extraction, pattern recognition, and database preparation. Most of studies toward generic object recognition focuses on feature extraction and pattern recognition.

Environments around object and angles of view are varied for each, even if the same object is taken in an image. In order to solve this problem, a huge training image database for various objects must be prepared. The existing studies often use manually-

prepared image database for the research as training database. However, it is hard work to prepare an appropriate image database in general because of the following three reasons: various concepts should be covered, huge number of images should be prepared for each concept, and appropriate labels should be given to the images. To tackle these problems, *Web image mining* has recently been reported [4,5]. Each individual presently uploads his/her photos on social Web service. Web Image mining realizes to obtain a huge image dataset from social Web service such as Flickr*[6]. These studies about Web image mining directly use the tags given to images on the Web service as the labels. The tags are generally given by multiple Web users including the poster, thus it might be appropriate as a general intuition. However, inappropriate tags are sometimes given, e.g., coined terms and meaningless strings. The existing studies remove these noisy tags based on statistic information for tags and images; semantics and visual features are not taken in the consideration.

This paper proposes an approach for generating hierarchical image database from Web image sharing service, that is huge and semantically stratified. The proposed approach consists of ontology and image clustering with image features in a hybrid. Ontology verifies the tags given to images on Flickr as reasonable, and constructs the hierarchical structure. And, the image database becomes conceptual, e.g., "Animal" is the hypernym of "Dog" and "Cat." Moreover, images for each tag are clustered based on Gaussian Mixture Model (GMM) with image features. Then more appropriate images for the tag are detected; noisy images are estimated and removed from the database.

2. Related Work

Some studies have generated a huge image database from Web [7]: that is, *Web image mining*. Noise images are included in the dataset automatically collected from Web. Thus, it is inefficient for learning to directly use such a database. From this fact, the existing studies subjectively prepare about 10 keywords. It is reported that the effectiveness of gathering images from Web increases by using the subjectively-prepared keywords as queries. Fergus *et al*. achieved 58.9% precision in gathering images from Web for the given 10 keywords [5]. However, the keywords are prepared by humans, and the keywords are directly used as labels for images; the semantics and appropriateness of the labels are not commented on.

The recent studies used general labels from knowledge such as *WordNet* [8] for Web image mining. Torralba has hierarchically mapped 80 million images from Web based on *WordNet* [9]. The Torralba's method does not remove noise images, however, showed relatively high accuracy on generic object recognition. From the Torralba's study, it was confirmed that hierarchic image database was effective for generic object recognition. Deng *et al*. proposed *Imagenet* [10], in which images obtained from Flickr were hierarchically structured based on WordNet. Since noise images were manually removed from *Imagenet*, so it was too costly to structure and expand the dataset. These studies structured hierarchic image dataset using knowledge. However, it has been remained on using concepts in knowledge as queries. It is impossible to find Web image mining while removing noise tags and images using both semantic relation in knowledge and visual features.

3. The proposed approach

For constructing an effective image database, both "semantics" and "appropriateness of images" should be covered. To realize these tasks, the proposed approach consists of two differential techniques in a hybrid: Web intelligence and pattern classification.

The procedures of the proposed approach are shown as the follows;

1. Images on Flickr are searched with the tags in an ontology as queries.
2. Effective tags for structuring hierarchy are selected with an ontology.
3. Noise images are removed with GMM-based clustering for each tag.

*http://flickr.com/

The proposed approach removes noise tags based on semantic relation between queries used for searching images on Flickr and the tags given to the searched images; these mean the above 1) and 2). And then, noise images are removed using image clustering with visual features for each tag; this means the above 3). Detail on the procedures will be described below.

3.1. Hierarchically-clustering images corresponding tags to ontology

This subsection details the step 1) and 2) in the proposed approach. The details about these steps have been described in our previous work [11]. Thus, this paper briefly explains the general description.

This paper uses Flickr as a source of images. Images on Flickr can be searched with given queries. The poster gives some tags to an image to suggest contents of the image. Images are searched with a query that equals class label on an ontology, i.e. *search class*. Ontology is knowledge structure which consists of concepts in a real world. In this paper, *DBpedia* [12] is used as an ontology, which is automatically generated from Wikipedia and covers varied concepts. Using the ontology, relation among concept (e.g., 'dog' is a 'mammal') can be recognized. When a concept A is a hyponym of another concept B, it is founded that "*A is a B*;" that is called as "IS-A relation."

The searched images have various and many tags. However, the tags are sometimes inappropriate for the image, e.g., meaningless strings and coined terms. These tags may be noise to structure conceptual hierarchy, and should be removed based on the semantics. At first, tags are verified with concepts in the ontology. The tags that does not correspond to the concepts are removed because it is likely to be coined terms. Introducing the idea of IS-A relation, it is determined whether the remained noise tags are removed or not. If a tag does not equal the search class or does not have IS-A relation with the search class, the tag is removed as noise. For example, in case of an image that is searched with a query "animal" has tags "animal," "dog," and "bazooka," then both "animal" and "dog" are remained as appropriate tags; "animal" equals to the search class, and "dog" has a IS-A relation with the search class "animal." Then, images are mapped on hierarchy (i.e., an ontology) according to the remained tags while the remained tags are used as labels for images. However, Word-sense and personal-name disambiguation are not resolved, these would be our future work.

3.2. Noise images removal by GMM-based image clustering

The images for each tag should be verified whether the contents of the image are appropriate for the label or not. The details about image features and the procedures to remove noise images will be shown below. This paper uses Fisher Vector [13] as image features, and removes noise images based on GMM-based image clustering.

For nth image, V-tuple SIFT features [14], which is shown as $\boldsymbol{X}_n = \{\boldsymbol{x}_n^v, v = 1,2,\cdots,V\}$, is calculated with grid sampling. Where, V is determined by interval and scope for extracting features, and the image size. As N shows the number of images in the database, $V * N$ SIFT features: $\boldsymbol{x}_n^v$ are calculated from the database. As clustering all $\boldsymbol{x}_n^v$ in the database into K classes with GMM clustering, the centroid: $\boldsymbol{\mu}_K$, the variance: $\boldsymbol{\sigma}_K$, and the mixing coefficient: ω_K are obtained, where K means the number of Visual Word for Bag of Features. Then, the obtained parameters of GMM are used as the Visual Word $\boldsymbol{\theta}$ of Fisher Vector as the follows;

$$\begin{aligned}\boldsymbol{\theta} &= (\omega_2,\cdots,\omega_K,\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_K,\boldsymbol{\sigma}_1,\cdots,\boldsymbol{\sigma}_K)\\ &= (\theta_1,\cdots,\theta_A), \qquad (1)\\ A &= (2b+1)K-1,\end{aligned}$$

where b shows the dimension number of SIFT features, and each $\boldsymbol{\sigma}_K$ and $\boldsymbol{\mu}_K$ is respectively calculated for each SIFT dimension.

From V-tuple $\boldsymbol{X}_n$ for an image, score function: $\boldsymbol{s}(\boldsymbol{X}_n|\boldsymbol{\theta})$ is calculated with reference to gradient vector for $\boldsymbol{\theta}$ as the following equation;

$$\boldsymbol{s}(\boldsymbol{X}_n|\boldsymbol{\theta}) = \left(\frac{\partial \log p(\boldsymbol{X}_n|\boldsymbol{\theta})}{\partial \theta_1},\cdots,\frac{\partial \log p(\boldsymbol{X}_n|\boldsymbol{\theta})}{\partial \theta_A}\right)^V. \quad (2)$$

Table 1. Statistical information about the generated hierarchic image database for labels that have more than 100 images.

Number of images whose label has more than 100 images	22,882
Number of terminal labels in which more than 100 images were gathered	84
Number of all labels in which more than 100 images were gathered (including hypernyms of terminal label as labels)	152
Average number of images for each terminal label	272

Fisher vector concerning nth image: $\boldsymbol{g}_n$ is obtained with $\boldsymbol{s}(\boldsymbol{X}_n|\boldsymbol{\theta})$ as the follows;

$$\boldsymbol{g}_n = \frac{1}{V}\boldsymbol{L}_{\boldsymbol{\theta}}\boldsymbol{s}(\boldsymbol{X}_n|\boldsymbol{\theta}). \tag{3}$$

Applying L2 norm and power norm to $\boldsymbol{g}_n$, $\boldsymbol{G}_n$ is calculated and assumed as image features for nth image.

As clustering images for a label l into C classes, the mean of each class: $\boldsymbol{m}_c^l$ is obtained as the following equation. Then, let I_c^l be the number of images for class c whose label is l, and images for a class c is represented as $\{\boldsymbol{G}_{c,i}^l, i = 1,\ldots,I_c^l\}$.

$$\boldsymbol{m}_c^l = \frac{1}{I_c^l}\sum_{i=1}^{I_c^l}\boldsymbol{G}_{c,i}^l, \quad c = 1,\cdots,C. \tag{4}$$

And, the mean vector for $\boldsymbol{G}_i^l$ concerning all images for label l: $\boldsymbol{M}^l$ is shown as follows;

$$\boldsymbol{M}^l = \frac{1}{I^l}\sum_{i=1}^{I^l}\boldsymbol{G}_i^l. \tag{5}$$

The Euclidean distance between each $\boldsymbol{m}_c^l$ and $\boldsymbol{M}^l$, that is d_c^l, is calculated as the following equation;

$$d_c^l = \sqrt{\sum_{j=1}^{A}(M_j^l - m_{c,j}^l)^2}, \tag{6}$$

where j means the index of vector dimension in $\boldsymbol{\theta}$.

Whether the images in class c are correct or noise for a label l is determined as follows;

$$\boldsymbol{G}_c^l = \begin{cases} \text{Correct images} & (d_c <= \text{T}), \\ \text{Noise images} & (d_c > \text{T}), \end{cases} \tag{7}$$

where, T shows the threshold. Here, the threshold T is defined as follows;

$$T = \frac{1}{C}\sum_{c=1}^{C} d_c. \tag{8}$$

If the class has large Euclidean distance to the average of d_c, images in the class is more likely to be noise for the label. That is, minority images for a label l are removed as noise images with this threshold.

4. Experiments to Generate Hierarchic Image Database

As of the experiments, 518 classes in DBpedia, which was an ontology used in this paper, were prepared. Less than or comparable to 4,000 images were obtained from Flickr for each query in a single search. SIFT features were extracted from each local point while window size for extracting local features was 16px and extraction interval was 8px. An image was expressed as Fisher Vector with number of visual word: $K = 512$. Only images whose label had more than 100 images were classified as the experimental setting.

4.1. Generated image database

TABLE 1 shows the statistical information about the generated hierarchic image database for only the labels that had more than 100 images. Fig. 1 shows the hierarchic structure of the generated image database, only some parts of the structure are expanded. It was confirmed that the generated image database had semantic hierarchy, e.g., $Thing - Species - Eukaryote - Animal - Mammal$.

Fig. 3 shows examples of images in the generated database. It was suggested that generic and appropriate images corresponding to label were obtained from Web without human. Especially, for even 'army' and 'fencing' which were abstract and not objects but matters, images were appropriately gathered in the database.

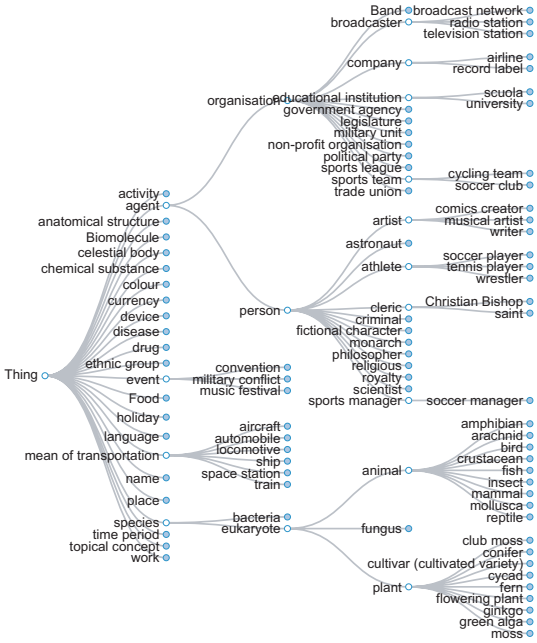


Fig. 1. Hierarchic structure of the generated image database.

4.2. Discussions about noise removal along with the steps in the proposed approach

The progress of the image database generation is discussed along with the steps in the proposed approach: removing noise tags with IS-A relation on the ontology, and removing noise images by using GMM-based image clustering.

TABLE 2 shows the transition of statistical information about the tags for each procedure. As shown in TABLE 2, only 835,598 out of 1,603,542 images had tags corresponding to concept on an ontology: that means about half of all images had only noise tag. Through removal of noise tags with IS-A relation on the ontology, the dataset finally had 47,910 images. Fig. 2 shows examples of images for 'cat' at this point. In the figure, images about not only 'cat' but 'tiger' and 'lion' were included in the image set whose label was 'cat.' 'Tiger' and 'lion' were semantically similar to 'cat,' therefore these noise tags remained slipping through the tag filtering with the ontology. These noise images would be removed by GMM-based image clustering, whose progress will be shown in detail below.

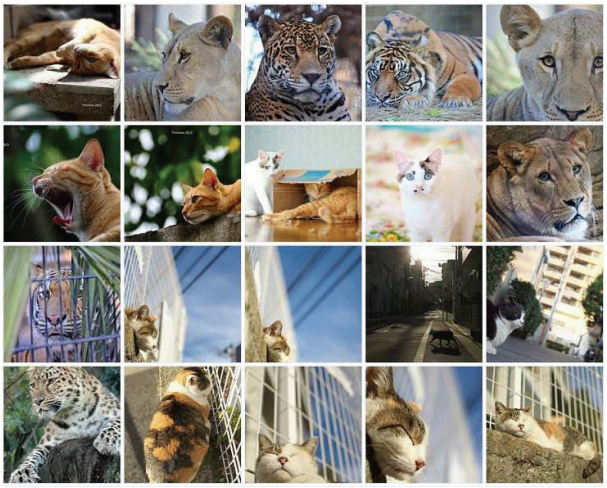
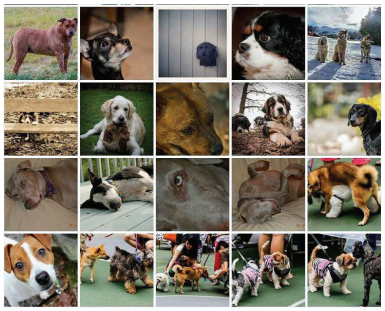


Fig. 2. Examples of images for 'cat' before step 3) in the proposed approach.

Through GMM-based image clustering, 7,259 out of 30141 images were detected as noise image and removed. Finally, as shown in TABLE 1, 22,882 images were gathered in the database as appropriate images. Figure 4 shows examples of appropriate and noise images for 'cat' and 'carb' detected by the proposed approach. As shown in the upper part of the figure, it was confirmed that most images about 'lion' and 'tiger' were removed as noise images for 'cat,' though few noise images are remained. In the



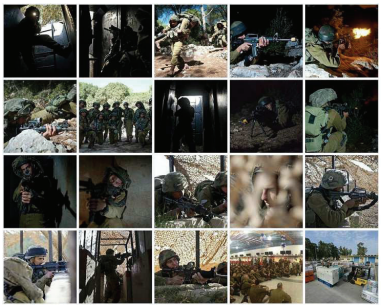
‘dog’

‘crab’

‘elephant’

‘man’

‘fencing’



‘army’

‘lizard’

‘spider’

Figure 3: Examples of images in the generated database: images for ‘dog,’ ‘crab,’ ‘elephant,’ ‘man,’ ‘fencing,’ ‘army,’ ‘lizard’ and ‘spider.’

Table 2. Transition of statistical information about tags for each procedure.

	Number of images	Number of the kinds of tags	Average number of tags for an image
Originally from Flickr	1,603,542	612,727	15.01
Tags are verified with an ontology	835,598	30,250	3.85
Noise tags are remove with IS-A relation on an ontology	47,910	4,872	1.55

similar manner, as shown in the lower and in the center of the figure, noise images for 'crab' were removed; these images were about meals using crabs as ingredients. However, images in the lower and in the right of the figure were wrongly removed as noise images for 'crab.' It could be said that the both correct and wrong detection for 'crab' should come from a same reason: the background color. The proposed approach considers not only object but also background as visual features for image clustering. The wrong detection was viewed as a negative effect of this characteristic.

5. Evaluation of the generated hierarchic image database

The image database evaluation experiments were conducted. The relationship between contents of image and the label was subjectively evaluated by humans. In the experiments, 13 kinds of labels were selected. 25 participants evaluated images for the each selected label whether the image was appropriate for the label or not. The image, that over 13 participants evaluated as appropriate for the label, was assumed as a correct image for the label. Then, precision is calculated as follows;

$$\text{Precision} = \frac{Positive}{All}, \tag{9}$$

where, $Positive$ and All each means the number of correct images for a label and the number of all images that the proposed approach detected as appropriate for the label, respectively.

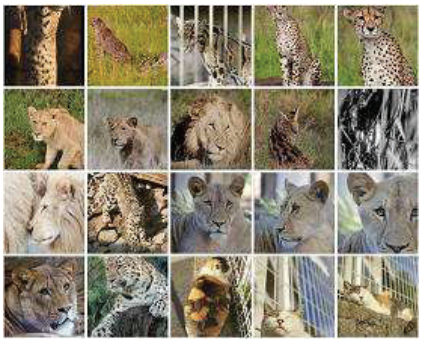
TABLE 3 shows the results of the evaluation experiments, where the precisions before step 3) are also shown at the second column for reference. From the third column of the table, it was confirmed that over 80% precisions for most labels. It was reasonable to say that the image database generated by the proposed approach was sufficiently usable as learning database for generic image recognition. However, for 'earth' and 'green,' the precisions were significantly low in compared with other labels. One of the reason of this might be that these labels were quite abstract and unexpectedly corresponded to varied things. For example, images for earth as a planet, garbage, and nature such as a forest were labeled as 'earth' in the database.

In order to confirm the effectiveness of step 3) in the proposed approach, that is GMM-based image clustering for removing noise images, let us focus on the second column of TABLE 3. Comparing the second column of TABLE 3 with the third of the one, it was confirmed that the precisions increased for most labels. These facts suggested that removing noise images was effectively functioned for high precision.

6. Conclusions

This paper proposed an approach to automatically generate hierarchic image database using an ontology and GMM-based image clustering in a hybrid. Using an ontology, the proposed approach automatically determined queries for searching images on Web, which the existing studies have prepared by heuristics. Then, wrong and unusable tags given to images were removed by referring IS-A relation in an ontology. Moreover, images for each label were verified whether the image was noise or not with GMM-based image clustering. That is to say, the proposed approach consists of techniques on two

Images estimated as appropriate for 'cat'



Images estimated as noise for 'cat' (1)

Images estimated as noise for 'cat' (2)

Images estimated as appropriate for 'crab'

Images estimated as noise for 'crab' (1)

Images estimated as noise for 'crab' (2)

Figure 4: Examples of images estimated as appropriate and noise for 'cat' and 'carb.'

differential fileds in a hybrid; Web intelligence and image processing.

The results of the evaluation experiments showed over 80% precisions as the relation between image contents and the label was appropriate. This fact suggested that the hierarchic image database generated by the proposed approach can be an image database for generic image recognition. Through the discussions about each step in the proposed approach, it was confirmed that the each process effectively functioned.

To evaluate actual effectiveness of the generated image database in the filed of image processing, the database will be applied to generic image recognition in the future. The effectiveness of the generated image database will be compared with the other image database, e.g., Caltech-256 [15] and PASCAL VOC [16]. Also, Word-sense and personal-name disambiguation in the verification of tags will be our future work.

Acknowledgments

Iwahori's research is supported by JSPS Grant-in-Aid for Scientific Research (C)(26330210) and Chubu University Grant. Yamanishi's research is supported in part by Artificial Intelligence Research Promotion Foundation and Ritsumeikan University Grant. The authors are thankful to the Lab related member of Chubu University and Ritsumeikan University.

References

1. R. Bergevin and M. D. Levine, "Generic object recognition: Building and matching coarse descriptions from line drawings," *Pattern Analysis and Machine*

Table 3. The results of the evaluation experiments.

Label	The precision before step 3)	Precision
cat	84.9	90.7
coffee	59.5	71.9
crab	90.8	96.7
cypraeidae	99.4	99.1
dog	98.2	97.9
earth	32.0	30.4
fencing	92.1	92.5
fern	77.7	82.9
fungus	93.5	96.2
green	45.5	46.9
lion	97.0	97.7
lizard	97.4	97.5
squirrel	94.3	96.2
AVG	81.7	84.4

Intelligence, IEEE Transactions on, vol. 15, no. 1, pp. 19–36, 1993.
2. Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. II–97.
3. A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 3, pp. 416–431, 2006.
4. T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *ACM International Conference on Image and Video Retrieval*, 2009, p. Article No.48 (9 pages).
5. R. Fergus, P. Perona, and A. Zisserman, "A Visual Category Filter for Google Images," in *Computer Vision-ECCV 2004*, 2004, pp. 242–256.
6. X. Li, C. G. Snoek, and M. Worring, "Unsupervised multi-feature tag relevance learning for social image retrieval," in *ACM International Conference on Image and Video Retrieval*, 2010, pp. 10–17.
7. X. Song, C.-Y. Lin, and M.-T. Sun, "Autonomous visual model building based on image crawling through internet search engines," in *the 6th ACM SIGMM international workshop on Multimedia information retrieval*, 2004, pp. 315–322.
8. G. A. Miller, "Wordnet: A lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
9. A. Torralba, "80 million tiny images : A large dataset for non-parametric object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelltigence*, vol. 30, no. 11, pp. 1958–1970, 2008.
10. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A Large-Scale Hierarchical Image Database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
11. R. Fujimoto, R. Yamanishi, Y. Iwahori, K. Toshioka, and J. Fukumoto, "Generation of stratified image database with web image sharing service and ontology," in *Proc. of IIAI 3rd International Conference on Advanced Applied Informatics*, 2014, pp. 966–971.
12. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web Journal*, 2013.
13. F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *IEEE Conference on Computer Vision and Pattern Recognition,*, 2007, pp. 1–8.
14. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
15. G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep. 7694, 2007. [Online]. Available: http://authors.library.caltech.edu/7694
16. M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.