An Analysis on Visual Recognizability of Onomatopoeia Using Web Images and DCNN features

Wataru Shimoda Keiji Yanai

Department of Informatics, The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585 JAPAN
{shimoda-k,yanai}@mm.inf.uec.ac.jp

Abstract. In this paper, we examine the relation between onomatopoeia and images using a large number of images over the Web. The objective of this paper is to examine if the images corresponding to Japanese onomatopoeia words which express the feeling of visual appearance can be recognized by the state-of-the-art visual recognition methods. In our work, first, we collect the images corresponding to onomatopoeia words using an Web image search engine, and then we filter out noise images to obtain clean dataset with automatic image re-ranking method. Next, we analyze recognizability of various kinds of onomatopoeia images by using improved Fisher vector (IFV) and deep convolutional neural network (DCNN) features. By the experiments, it has been shown that the DCNN features extracted from the layer 5 of Overfeat's network pre-trained with the ILSVRC 2013 data have prominent ability to represent onomatopoeia images.

Keywords: onomatopoeia, Web images, DCNN features

1 Introduction

In general, an "onomatopoeia" is a word that phonetically imitates, resembles or suggests the source of the sound that it describes such as "tic tac" and "quack". In English language, an onomatopoeia is commonly used only for expressing sounds in everyday life. However, onomatopoeia words in Japanese language are commonly used in the boarder purpose such as expressing feeling of visual appearance or touch of objects or materials. Figure 1 shows a "fuwa-fuwa" object, which means being very softy like very soft cotton. In Japanese language, there are so many onomatopoeia words like "fuwa-fuwa" expressing some kinds of feeling of appearance or touch.

The relation between images and onomatopoeia has not been never explored in the context of multimedia research, although many works related to words and images have been done so far. Then, in this paper, we try to analyze the relation between images and onomatopoeia by using a large number of tagged images on the Web. Especially, we examine if onomatopoeia images can be recognized by the state-of-the-art visual recognition method. As a case study on onomatopoeia images, we focus on onomatopoeia in Japanese language, because Japanese language has much more onomatopoeia words which are used in the more broader context compared to other languages such as English.

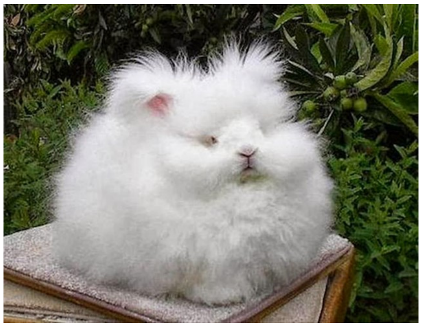

Fig. 1. An example photo of “Fuwa-fuwa” object.

In this paper, we collect images corresponding to Japanese onomatopoeia words representing feeling of appearance or touch of objects from the Web, and then analyze the relation between onomatopoeia words and images corresponding to them in terms of recognizability using two kinds of state-of-the-art image representations, Improved Fisher Vector [6] and Deep Convolutional Neural Network Features (DCNN features) [8].

2 Related Works

In this section, we mention some works on material recognition as related works on onomatopoeia.

Since Japanese onomatopoeia represents feeling of appearance, recognition of onomatopoeia image is more related to material recognition than generic object recognition. As works on material recognition, the work on Flickr Material Database (FMD) [5] is the most representative. They constructed FMD which consists of ten kinds of material photos, “Fabric”, “Foliage”, “Glass”, “Leather”, “Metal, “Paper”, “Plastic”, “Stone”, “Water” and “Wood”. Each of these material classes has unique visual characteristics which enables people to estimate which material class a given material photo belongs to. However, it was unexplored what kinds of visual features are effective for it. The situation was different from object recognition where local features and bag-of-features representation were proved to be effective. Liu et al. [5] proposed a method to classify material photos based on topic modeling with various kinds of image features. They achieved 44.6% classification accuracy. Cimpoi et al. [1] proposed to represent material images with state-of-the-art image representations, Improved Fisher Vector [6] and Deep Convolutional Neural Network Features (DCNN features) extracted by DeCAF [2], and achieved 67.1% for 10 class material photo classification of FMD. They also created the larger-scale textured photo database, Describable Textures Dataset (DTD), which consists of 47 classes as shown in Figure 2, and proposed to use them as texture attributes. Inspired by their work, we also use IFV and DCNN features in this paper.

Both FMD [5] and DTD [1] are constructed by gathering images from the Web and selecting good images by hand. Since DTD is relatively a large-scale dataset, they used crowd-sourcing service, Amazon Mechanical Turk (AMT), to select good images out of the images gathered from the Web. Nowadays, AMT is commonly used to image filtering. However, it costs more than a little expense. In this work, we adopt fully automatic image gathering method to built an onomatopoeia image dataset based on the method on automatic Web image gathering and re-ranking with pseudo-positive training samples [9, 7]. An automatic method is helpful to prevent human's prejudice from getting into the process of image selection.

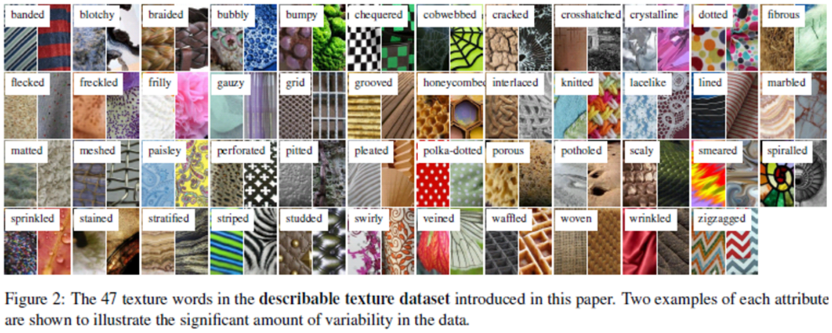

Figure 2: The 47 texture words in the **describable texture dataset** introduced in this paper. Two examples of each attribute are shown to illustrate the significant amount of variability in the data.

Fig. 2. 47 categories in the DTD dataset (cited from [1]).

3 Methods

In this paper, first we construct an onomatopoeia image database automatically, and next analyze the relation between onomatopoeia words and the corresponding images in terms of visual recognizability of onomatopoeia words.

3.1 Gathering onomatopoeia images

To gather onomatopoeia image, we use Bing Image Search API by providing Japanese onomatopoeia words as query words. Most of the upper-ranked images in the search results can be regarded as the images which correspond to the given onomatopoeia word. However, some images irrelevant to the given word are expected to be included even in the upper-ranked results. Therefore, we re-rank the results obtained from Bing Image Search API so that only relevant images are ranked in the upper rank. To re-rank images, we use the similar approach as [7, 9] where no human supervision is needed. We regard the upper-ranked images in the search result as pseudo-positive training samples and random images as negative samples, and train SVM with them. Then, we apply the trained SVM to the images in the original search results, and sort images in the descending order of the SVM output values to obtain re-ranked results. In our work, we repeat this re-ranking process twice. The detail of the procedure of image collection is as follows:

(1) Prepare Japanese onomatopoeia words.

(2) Gather 1000 images corresponding to each onomatopoeia word using Bing Image Search API.

(3) Extract an image feature vector from each of the gathered images using Improved Fisher Vector [6] and Deep Convolutional Neural Network Features (DCNN features) [8].

(4) Regard the top-10 images in the search result as pseudo-positive samples and random images as negative samples, and train a linear SVM with them.

(5) Apply the trained SVM to the images in the original search results, and sort images in the descending order of the SVM output values.

(6) Carry out the second re-ranking step. Train a linear SVM with the top-20 images in the re-ranked results as pseudo-positive samples, apply it, and sort images in the descending order of the SVM output values again.

(7) Finally regard the top-50 images as the images corresponding to the given onomatopoeia word.

3.2 Evaluation of recognizability of onomatopoeia words

After gathering onomatopoeia images, we evaluate to what extent the images corresponding to an onomatopoeia word can be recognized by state-of-the-art object recognition methods.

We mix 50 onomatopoeia images and 5000 random noise images and discriminate onomatopoeia images from noise images, and examine if we can separate onomatopoeia images from noise images for these 5050 mixed images by visual recognition methods regarding each of the onomatopoeia image sets.

To classify onomatopoeia images, we regard 50 onomatopoeia images selected in the previous step as positive samples and 5000 random images as negative samples, and train a linear SVM. Then, we apply the trained SVM into the mixed image set containing 5050 images and rank all the images in the descending order of the SVM output values, and evaluate the result with average precision. In our work, we regard that the obtained average precision means the recognizability of the corresponding onomatopoeia word.

The average precision is calculated in the following equation:

$$AP = \frac{1}{m}\sum_{k=1}^{m} Precision_{true}(k)$$

, where m is the number of positive sample (50), and $Precision_{true}(k)$ means the precision value within the k-th positive samples.

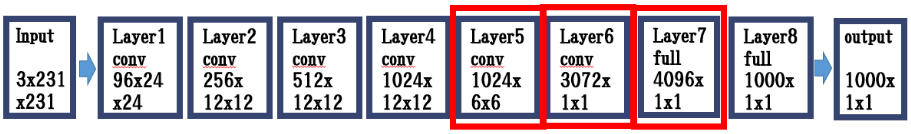

Fig. 3. The structure of the Deep Convolutional Neural Network (DCNN) for the ILSVRC 2013 dataset in Over feat[8]. We extracted feature vectors from the Layer-5, the Layer-6 and the Layer-7.

3.3 Image Features

As image representation in both the image collection step and the evaluation step, we use two kinds of state-of-the-art features, Improved Fisher Vector [6] and Deep Convolutional Neural Network Features (DCNN features) [8].

Improved Fisher Vector（IFV） To encode an image to IFV, we follow the method proposed by Perronnin et al. [6]. First, we extract SIFT local features randomly from a given image, and apply PCA to reduce their dimension from 128 to 64. Next, we code them into a Fisher Vector with the GMM consisting of 64 Gaussian, and obtain IFV after L2-normalizing the Fisher Vector. Since the dimension of the IFV is $2DK$ where D is the number of dimension of the local features and K is the number of elements of MGM, totally it is $2 \times 64 \times 64 = 8192$.

3.4 Deep Convolutional Neural Network (DCNN)

Recently, it has been proved that Deep Convolutional Neural Network (DCNN) is very effective for large-scale object recognition. However, it needs a lot of training images. In fact, one of the reasons why DCNN won the Image Net Large-Scale Visual Recognition Challenge (ILSVRC) 2013 is that the ILSVRC dataset contains one thousand training images per category [4]. This situation does not fit common visual recognition tasks Then, to make the best use of DCNN for common image recognition tasks, Donahue et al. [2] proposed the pre-trained DCNN with the ILSVRC 1000-class dataset was used as a feature extractor.

Following Donahue et al. [2], we extract the network signals from the middle layers (layer 5, 6 and 7) in the pre-trained DCNN as a DCNN feature vector. We use the pre-trained deep convolutional neural network in Overfeat [8] as shown in Figure 3. This is slight modification of the network structure proposed by Krizhevsky et al. [4] at the LISVRC 2012 competition. In the experiments, we extract raw signals from layer-5, layer-6 or layer-7, where the dimension of the signals are 36864, 3072 and 4096, respectively, and L2-normalize them to use them as DCNN feature vectors.

3.5 Support Vector Machine（SVM)

For classification in both the image collection step and the evaluation step on recognizability of onomatopoeia images, we use a linear SVM which is commonly used as

Table 1. Twenty kinds of Japanese onomatopoeia words used in the experiments.

onomatopoeia	meaning	onomatopoeia	meaning
pika-pika	shining gold	mofu-mofu	softly
bash-basha	splashing water	mock-mock	volumes of smoke; mountainous clouds
fuwa-fuwa	softly; spongy	kara-kara	hanging many metals
nyoki-nyoki	shooting up one after another	bou-bou	overgrown
kira-kira	shining stars	fuwa-fuwa	well-roasted
gune-gune	winding	siwa-siwa	wrinkled; crumpled
toge-toge	thorny; prickly	zara-zara	sandy; gritty
butsu-butsu	a rash	kari-kari	crispy; crunch
puru-puru	fresh and juicy	guru-guru	whirling
gotsu-gotsu	rugged; angular; hard; stiff	giza-giza	notched; corrugated

a classifier for IFV and DCNN, since they are relatively higher dimensional. In the experiments, we used LIBLINEAR [3] as an implementation of SVM.

4 Experiments

In the experiments, we collected images related to twenty onomatopoeia words and examined their recognizability with Fisher Vector and DCNN features. The twenty Japanese onomatopoeia words we used in the experiments and their meanings are shown in Table 1, the visual recognizability of which we will examine in the experiments.

4.1 Data Collection

We gathered 1000 images for each of the twenty Japanese onomatopoeia words using Bing Image Search API, and repeated re-ranking twice using four kinds of image features. Finally we obtained an onomatopoeia image dataset containing twenty onomatopoeia categories where each category has fifty images without any human supervision. Figure 4 shows some images corresponding to ten onomatopoeia words.

We evaluated the precision of the onomatopoeia datasets constructed with four different kinds of image representations by subjective evaluation. Figure 5 shows the precision value of the selected fifty images on each of the twenty given onomatopoeia words in case of using IFV, DCNN Layer-7, DCNN Layer-6 and DCNN Layer-5 as a feature, respectively. As a results, DNN features outperformed IFV clearly.

4.2 Evaluation of recognizability

Figure 6 shows the results on recognizability of each of the twenty onomatopoeia words represented by the average precision of the results of separation of 50 onomatopoeia images from 5000 noise images in case of using IFV, DCNN Layer-7, DCNN Layer-6 and DCNN Layer-5 as a feature, respectively.

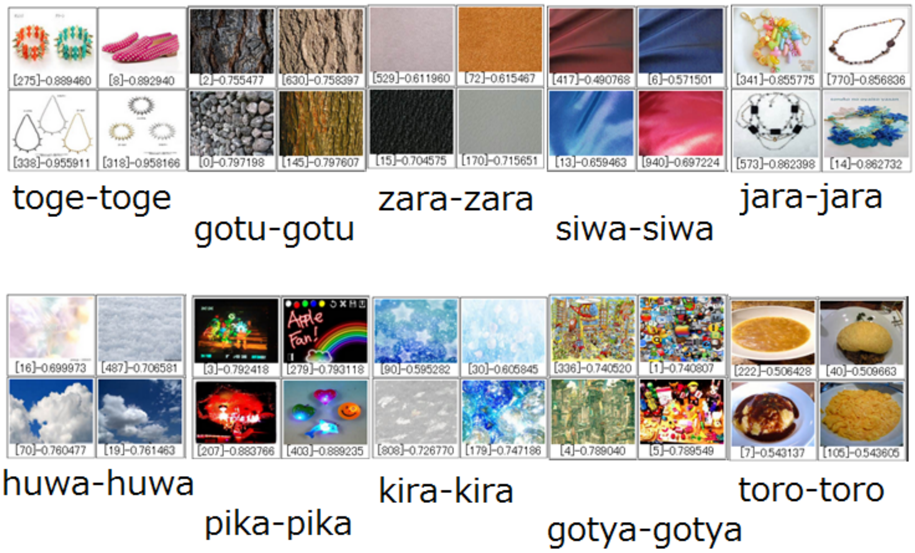

Fig. 4. Examples of onomatopoeia images gathered from the Web.

Figure 7, 8, 9 and 10 shows the top-20 "gotsu-gotsu" (which means being stiff or hard) images in the descending order of the SVM output values. In case of IFV, separation was failed, because the top-20 images contains many images irrelevant to "gotsu-gotsu". On the other hand, all the results by DCNN does not contain prominent irrelevant images. This result shows that DCNN features has high ability to express visual onomatopoeia elements in images.

5 Conclusions

In this paper, we examined if the images corresponding to Japanese onomatopoeia words which express the feeling of visual appearance or touch of objects can be recognized by the state-of-the-art visual recognition methods. In our work, first, we collect the images corresponding to onomatopoeia words using an Web image search engine, and then we filter out noise images to obtain clean dataset with automatic image re-ranking method. Next, we analyze recognizability of various kinds of onomatopoeia images by using improved Fisher vector (IFV) and deep convolutional neural network (DCNN) features.

By the experiments, it has been shown that the DCNN features extracted from the layer 5 of Overfeat's network pre-trained with the ILSVRC 2013 data have prominent ability to represent onomatopoeia images.

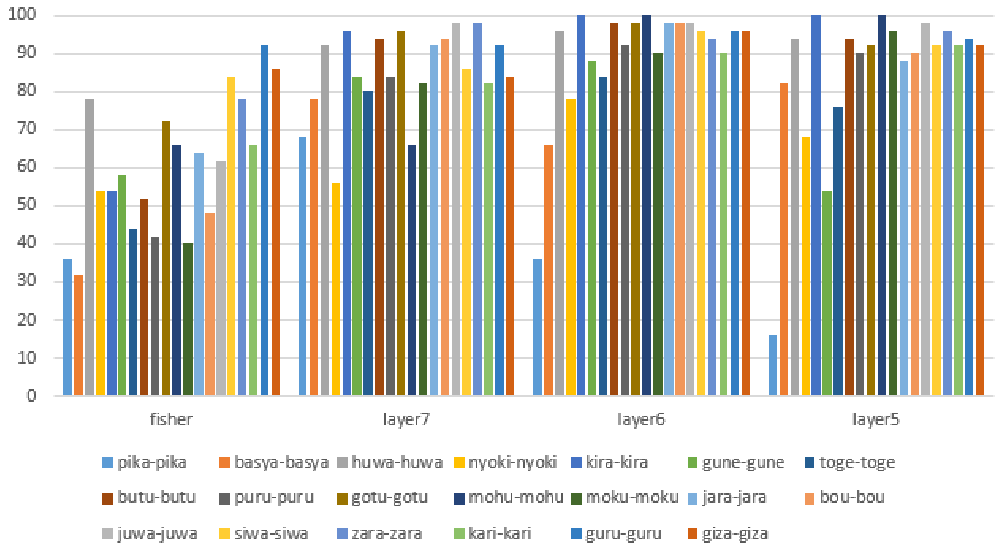

Fig. 5. Precision of the collected images corresponding to the 20 given onomatopoeia words.

References

1. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proc. of IEEE Computer Vision and Pattern Recognition (2014)
2. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531 (2013)
3. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research 9, 1871–1874 (2008)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
5. Liu, C., Sharan, L., Adelson, E., Rosenholtz, R.: Exploring features in a bayesian framework for material recognition. In: Proc. of IEEE Computer Vision and Pattern Recognition (2010)
6. Perronnin, F., Sanchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Proc. of European Conference on Computer Vision (2010)
7. Schroff, F., Criminisi, A., Zisserman, A.: Harvesting image databases from the web. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(4), 754–766 (2011)
8. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. In: Proc. of International Conference on Learning Representations (2014)
9. Yanai, K., Barnard, K.: Probabilistic Web image gathering. In: Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval. pp. 57–64 (2005)

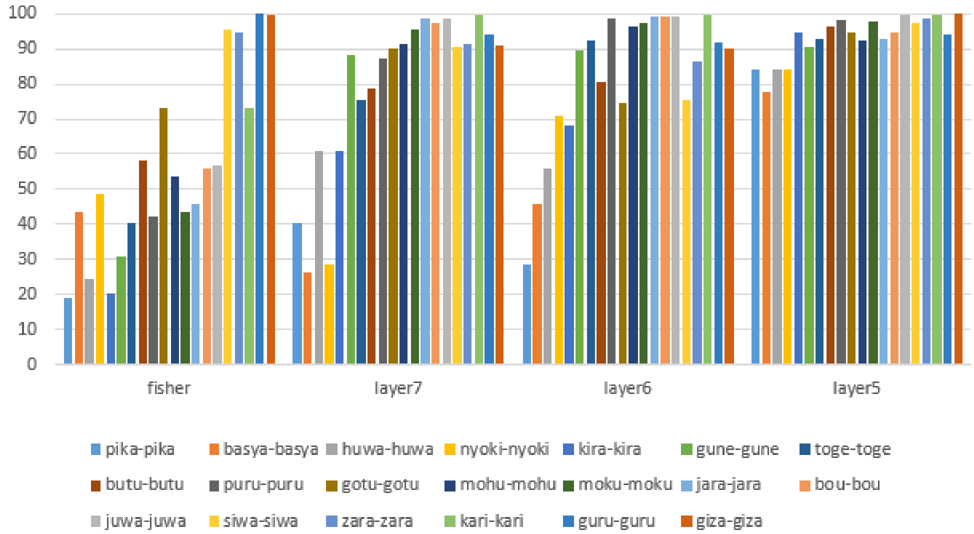

Fig. 6. Evaluation results of the recognizability of each onomatopoeia word.

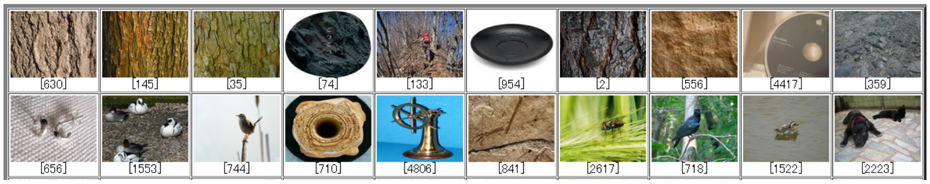

Fig. 7. The top-20 images of "gotsu-gotsu" classified with IFV features.

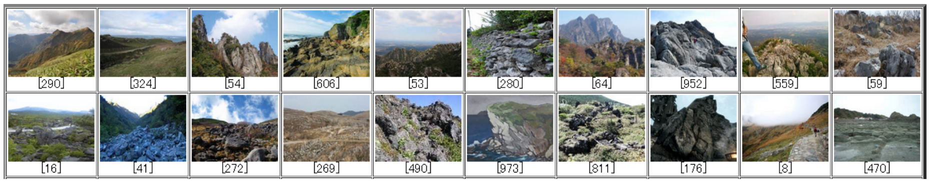

Fig. 8. The top-20 images of "gotsu-gotsu" classified with DCNN Layer-7 features.

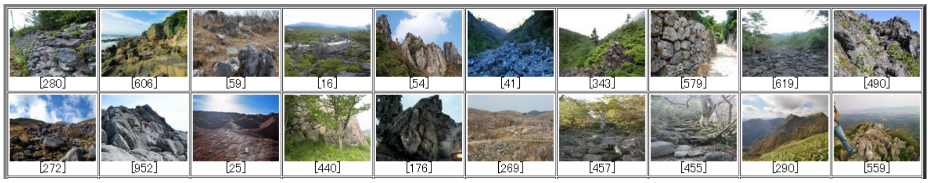

Fig. 9. The top-20 images of "gotsu-gotsu" classified with DCNN Layer-6 features.

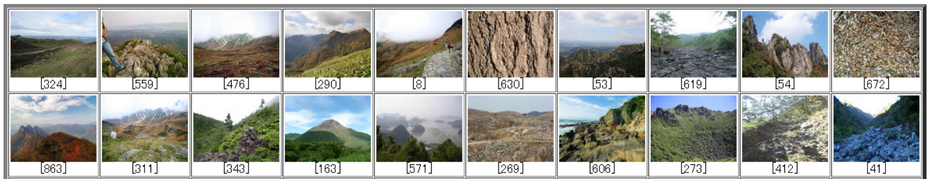

Fig. 10. The top-20 images of "gotsu-gotsu" classified with DCNN Layer-5 features.