AUTOMATIC IMAGE DESCRIPTION BY USING WORD-LEVEL FEATURES

Shingo Horiuchi
Service Innovation Center
Reseach and Development
Headquarters
NTT DATA, Japan
horiuchisng@nttdata.co.jp

Hirotaka Moriguchi
Department of Computer
Science
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, Japan
hmori@nii.ac.jp

Xu Shengbo
Department of Computer
Science
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, Japan
s-jyo@nii.ac.jp

Shinichi Honiden
Department of Computer
Science
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, Japan
honiden@nii.ac.jp

ABSTRACT

Automatic image description is one of the challenging tasks of image recognitions. However, there are image descriptions that contain some too specific phrases that cannot be judged only from appearance of images. In this paper, we propose a novel approach to collect general phrases for generating image descriptions. On the assumption that there are high frequency phrases related to an query image in the image descriptions of similar images, we select nouns and their attribute phrases from the image descriptions of similar images based on their frequency. In order to evaluate the relevance of our image description, we conduct comparative experiments with existing approaches. Our experimental results show that our image descriptions are short, concise and visually relevant to query images.

Categories and Subject Descriptors

I.4.0 [IMAGE PROCESSING AND COMPUTER VISION]: General—Image processing software

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Image Description, Image Recognition, Automatic Image Annotation.

ICIMCS'13, Aug. 17-19, 2013, Huangshan, Anhui, China.

1. INTRODUCTION

Understanding of semantics of these image data has important role to manage and search them efficiently. Many attempts have been done to understand the semantics of images. Image annotation is one of these approaches and it associates images with the set of words or short phrases related with the contents of the images. Manual annotation provides reliable tags but is not practical. However, manual annotation is time-consuming and it is unrealistic to put annotations on all images. Therefore, many researchers geared toward building automatic annotation systems.

In automatic image annotation, there are many approaches such as object recognition, scene recognition and image segmentation. The advance in the computer vision and the computational resources has enhanced the performance of these tasks rapidly. The performances of these tasks depend on the features extracted from images and these features become much more sophisticated in recent years.

Automatic image description is one of these automatic image annotation approaches and many researchers have been tackling it [2–6]. This task is difficult because this is a complicated combination of various computer vision tasks and natural language processing tasks. At the cost of this difficulty, the image descriptions have high-dimensional information of relationships between objects in the images and more detailed information about objects in the images. Some researchers have tackled the relationship extraction [1, 3].

In this paper, we propose an approach for automatic image description focusing on the more detailed information about object. There are some approaches to describe a query image using an image description of similar image or using subsets of words and phrases of image descriptions of similar images [2, 4, 5]. However, there are too specific words to attach to other images and there is information which cannot be judged only from the appearance of the images. As a result, too specific phrases turn into false description and phrases that are not judicable only from the appearance of images may be true, but may be false. In order to generate

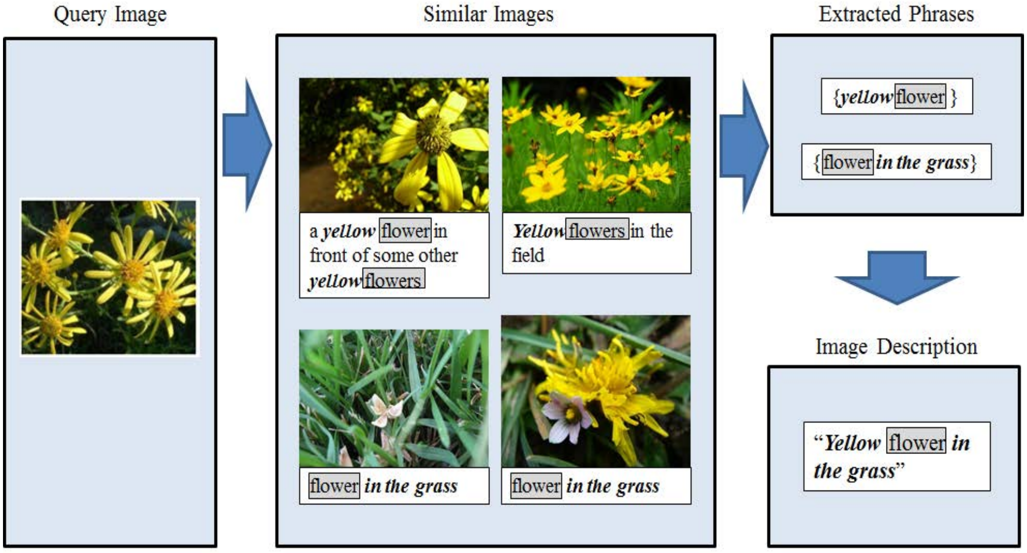

Figure 1: Pipeline of the proposal. First, we retrieve image descriptions of images which are similar to the query image. Second, we extract phrases in the image descriptions according to phrase frequencies and similarity ratios. Finally, we arrange the phrases to generate a image description of the query image.

visually relevant image descriptions, our proposed method contains three steps (Fig. 1). First, we retrieve image descriptions associated with the images similar to the query image. Second, we use the phrase frequencies and similarity ratios as the word-level features. Finally, we extract the detailed but general nouns and their attribute phrases based on word-level features. To arrange them in the form of sentences, we classify the attribute phrases into three types; adjective phrase, verbal phrase and prepositional phrase. Then the top scored attribute phrases are selected as the components of the image description. The image description is generated from these phrases by arranging them according to simple rules.

We conduct experiments to evaluate our image description. In these experiments, the descriptions generated by our method are compared with the descriptions generated by existing approaches and they are evaluated manually. Finally, we show that our approach generates image descriptions that contain fewer fault components than existing approaches.

The rest of this paper is organized as follows. Chapter 2 describes our proposal approach in detail. Chapter 3 explains about the experiments to evaluate our image descriptions and their results. Chapter 4 discusses image descriptions generated by our method and the results of experiments. Finally, chapter 5 shows our conclusion and open issues.

2. APPROACH

2.1 Similarity-based Image Retrieval

2.1.1 SBU Caption Photo Dataset

SBU Caption Photo Dataset is a set of 1 million captioned photographs [5]. The photographs were collected from Flickr and filtered in order to get visually relevant image descriptions associated with images. V. Ordonez et al. simply associate a query image with an image description for the other image; therefore they need image descriptions that describe visual contents of images. To collect such image descriptions, they select the images with image descriptions such that they contain long phrases describing visual contents, more than two words of their term list and at least one prepositional phrase.

2.1.2 Feature Detection

To collect image descriptions including the phrases related to the contents of a query image, we apply similarity-based image retrieval using the large scale dataset. Thus, it is desirable to use simple feature descriptors that are low-dimensional vectors. In this paper, we adopt GIST and TinyImages as image descriptors for similarity-based image retrieval [5].

GIST is a global image descriptor. TinyImages is a global color descriptor extracted from a scale-downed image whose size is 32 × 32.

2.1.3 Data Augmentation

Two descriptors mentioned in Section 2.1.2 are extracted only from squared image, thus a query image have to be transformed into square. Therefore all input images and images in the dataset must be transformed. There are two types of picture forms in general and almost all images belong to these two types: tall one and wide one. Therefore there are mainly two options to make images square. In our similarity-based image retrieval, only simple descriptors are used, thus this two types of transformation turn to big problem.

In the same way, these simple descriptors are sensitive to where the object and stuff in an image are. Therefore wide variety of image appearance is helpful to get similar images from dataset. Moreover horizontal reflection is also helpful to get similar images.

Finally, we get 12 reconstructed images from the original input image after all.

2.1.4 Collected Image Description

When a query image is given, this similarity-based image retrieval approach outputs the ranking of 1 million images in SBU Caption Photo Dataset sorted by similarity ratios between the query image and them.

The subsets of entire image descriptions are used in text processing approach mentioned in Section 2.2. Here, we adopt 1 thousand image descriptions of top 1 thousand images in this ranking for each reconstructed input images. Because we have 12 reconstructed images for a query image, we get 12 thousands of image descriptions for the query image. This subset of entire image descriptions is very noisy but has much visual information similar to the contents of a query image.

2.2 Text Processing

2.2.1 Word-level Feature

In our text processing approach, we select nouns from the image descriptions of similar images as object descriptions in the query image, and then select their attributes as descriptions for the objects and the scene of the entire image. The image descriptions collected by similarity-based image retrieval are very noisy, and thus we have to select relevant nouns and their attributes from image descriptions. To select relevant phrases such as nouns and attributes, we define word-level feature (WLF) as criteria based on phrase frequency.

$$\mathrm{WLF}(t) = (\log_2(\mathrm{frq}(t) \cdot \mathrm{iRn}(t)) + 1) \cdot \mathrm{ridf}(t) \qquad (1)$$

$$\mathrm{iRn}(t) = \sum_{s \in S'(t)} \frac{1}{\mathrm{sim}(s)} \qquad (2)$$

$$\mathrm{ridf}(t) = \frac{1}{\log_2(all\mathrm{frq}(t)) + 1} \qquad (3)$$

Where, $frq(t)$ is the frequency of phrase t in the collected image descriptions, $S'(t)$ is the subset of the collected image descriptions which includes phrase t, $sim(s)$ is the similarity ratio between the image associated with the image description s and the query image, and $allfrq(t)$ is frequency of phrase t in the entire 1 million image descriptions.

2.2.2 Noun Selection

In our approach, Nouns are selected at first.

Overview of this procedure is

After similarity-based image retrieval, we have collected image descriptions for the query image. Our descriptors used in similarity-based image retrieval are too simple to select proper nouns. Thus, we get help of class categories which indicate objects in the query image. Practically, such class categories are outputted as a result of object recognition that is one of the hot topics of the image processing. In order to evaluate the effectiveness of our automatic image description generating approach, we regard these class categories as given below Oracle. However, if the numbers of classes increase, the more classifier and training set are need. Our noun selecting procedure attacks this problem by using the tree structure of WordNet.

We filter all nouns in collected image descriptions by synsets of WordNet. In other words, the nouns such as the class categories themselves, hyponyms of them and hypernym of them are selected as candidates. After filtering nouns, these candidates are weighted according to the information contents of the synsets. At last, WLF values of these weighted candidates are calculated and the ranking based on WLF is outputted. The noun whose WLF weighted information contents is highest is selected as object description.

2.2.3 Attribute Selection

To get more information from collected image descriptions, attribute phrases of selected nouns are extracted. In our approach, we define 3 types of attribute phrases and select the most relevant one for each of them. These 3 types consist of a) adjective phrase, b) verbal phrase and c) prepositional phrase. Because the information of which type the attribute phrases are can be received as the output of parser, attribute classification can be automatically done.

Here, we introduce another assumption that too specific attribute phrases are not suitable for automatic image description for an unseen image. According to this assumption, too specific attribute phrases are omitted by checking regularized idf scores. The thresholds of omitting are set to 3 types of attribute phrases respectively by checking the regularized idf scores.

After selecting candidate attribute phrases, WLF values of them are calculated and the ranking based on WLF values are outputted respectively. If the WLF value is higher than the threshold, the attribute phrase is selected. If this value is lower than the threshold, it is judged that any attribute phrase in collected image descriptions is not suitable for the query image and no attribute phrase is selected. These thresholds are set to 3 types of attribute phrases respectively by pre-processing.

2.2.4 Description Generation

As output of our approach, the image description for the query image is generated by using selected nouns and their attribute phrases. Here, there are 3 types of attributes; adjective, verbal and prepositional phrases. By introducing very simple rules, the image description can be easily constructed because an image description does not always include all components of sentence. The rules are described below.

Where, n1 adj. is the selected adjective phrase of noun1 and n1 action is the selected verbal phrase of noun1. When there are more than two nouns, following nouns are arranged like the components of noun2. Prep. is the prepositional phrase of the sentence. If there is no component as adjective or verbal phrase, this component is skipped. If there is no prepositional phrase, image description generated by our method is too concise to be sentence. Therefore, we use general phrase “This is the picture of”. This is better than adding irrelevant prepositional phrase by selecting low WLF scored preposition.

3. EXPERIMENT AND RESULT

3.1 Experiment

In our experiment, we use 20 class categories used in PASCAL Visual Object Classes Challenge. In concrete terms, airplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, table(dining table), dog, horse, motorbike, person, plant + pot(potted plant), sheep, sofa, train, tv (tvmonitor) are used. Moreover, to compare with three images from the exiting approaches, we use additional four class categories; window, truck, flower and cloud.

We use two approaches as the baseline. One is the approach of [5] as 1 million and anothe is the approach of [4] as Integer Linear Programming (ILP).

We conduct two small experiments to detect the characteristics of our approach. Experiment a) uses 20 images randomly selected for the each class category. In this experiment, the results of our approach are compared with the imitation of 1 million approaches in order to check whether our approach can be applied to the images which have different object contents. Experiment b) uses 7 images which are in the paper [4] and the results of this are compared with ILP approach.

As the evaluation, we ask ten evaluators to review each component in generated image descriptions is true or not by judging from only visual information of a query image and then the numbers of true components and false components are counted, each image description is scored from 1 to 5 (5: perfect, 4: almost perfect, 3: 70-80% good, 2: 50-70% good, 1: totally bad) in the aspect of relevance. They are male Japanese graduated students of computer science.

3.2 Result

Examples of image descriptions of our approach and existing approaches are listed in Fig. 2. By comparing the outputs of our approach with the others, it is found that our image descriptions are short and consist of the general words.

Fig. 3 shows the average score for each query image per images in experiment(a).

From the results of experiment(a), the image descriptions generated by our method are more relevant than those generated by 1 million in almost all images. In the image of ship, horse and dog, our image descriptions are scored low. To find statistical significances, we conduct some Student's t-test. From this test, there are statistical significances at 0.05 in almost all images that are scored higher in our approach. With respect to the images scored lower in our approach, there is no statistical significance at 0.05.

Fig. 4 shows the average score for each query image per images in experiment(b).

From the result of experiment(b), in all except the image of bird (No. 3), the image descriptions generated by our method are scored high and we conduct some Student's t-tests as well as experiment(a). There are statistical significances at 5%. In the image of bird, there is no statistical significance at 5% (p-value = 0.425).

With respect to each evaluator, all evaluators score our image descriptions higher in average. In each score of evaluator, there is a statistical significance at 0.01.

4. DISCUSSION

Compared with the image descriptions generated by existing approaches, our image descriptions are short and consist of general words. The reason why our image descriptions are short is that at most one attribute phrases for each type is selected in our method. The reason why our image descriptions consist of the general words is that our approach selects the phrases whose frequencies are comparatively high. The high phrase frequencies mean that the phrases are used in common. Therefore, our method can omit too specific phrases by using the word-level feature. In experiment (a), 17 / 20 of our image descriptions are scored higher than the image descriptions selected by 1 million. This can be explained by saying that it is not suitable to attach image descriptions to the other images and too specific phrases (for example "the HOLLYWOOD sign", "UP pipe train" and "over 3rd floor roof") seldom are able to be used to explain other images. Our approach selects the general phrases to describe other images, therefore the suitable image description can be generated.

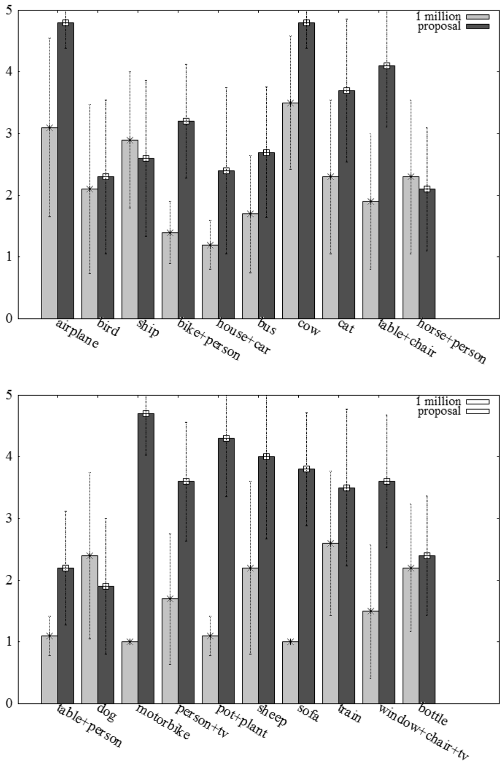

Figure 3: Average scores of 1 million and our approach per images in experiment(a).

In experiment (b), 6 / 7 of our image descriptions are scored higher than the image descriptions of ILP. This result shows that the attribute phrases cannot be used to explain other images even if the two images are similar in the feature spaces concerned with the attribute types. On the one hand, the image descriptions of ILP mention much about the information which is not in the query images (e.g. "Cat likes hanging around in my recliner" and "on first avenue") and cannot be judged only from the appearance of images (e.g. "east village", "in a rock garden" and "by river"), but on the other hand, our image description mention only about the contents of query images basically. Thanks to use of the word-level feature, our approach can select such phrases.

In overall experiments, all evaluators score our descriptions higher than the descriptions of existing approaches.

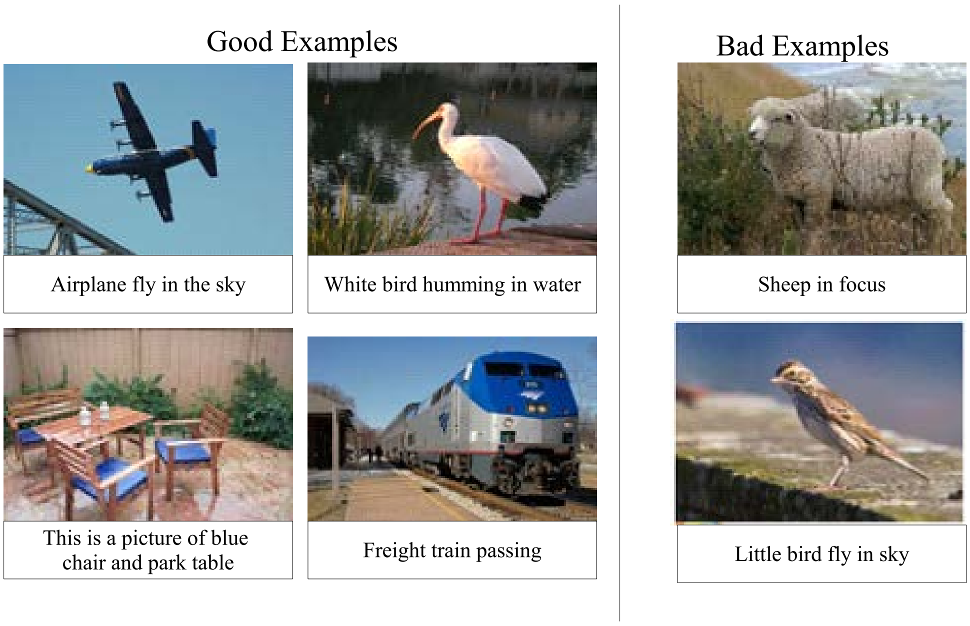

Figure 2: Good and Bad examples of our image descriptions.

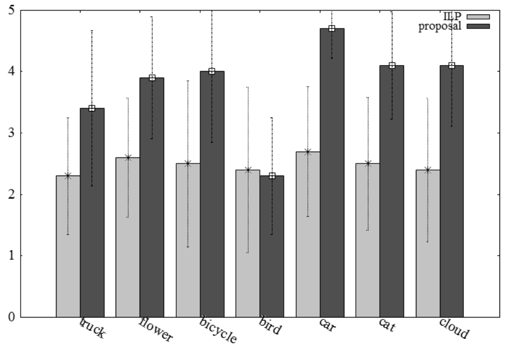

Figure 4: Average scores of ILP and our approach per images in experiment(b).

This means that, generally speaking, our descriptions contain less fault components and our descriptions are visually relevant to the query images.

We have shown the good point of our approach and explain about overall performance but, off course, there are some disadvantages in our image description.

First, some image descriptions are scored lower than those of existing approaches. This is mainly because we only use gist and tiny images description for similarity image retrieval and we can collect only roughly similar images as a result.

In addition, our approach cannot describe more detail about low frequent objects (e.g., motorbike, tv). In our experiment, the word frequency of “motorbike” in the descriptions of similar images is 1 and no attribute phrase is added for example. To put it better, our approach select not to add irrelevant phrases, however this is the obvious limitation of our approach. This is partly because the dataset and partly because our approach. If the dataset contains much more image descriptions, if the more detailed features are used in the similarity image retrieval and if attribute phrases related with other nouns can be selected, probably, this limitation is solved.

5. CONCLUSION

We proposed a novel approach for image description with few errors detecting the objects of the image contents and selecting their attribute phrases by using the word-level features from annotated images. In this approach, we introduced the novel criteria as the word-level features to select the most relevant nouns and their attribute phrases according to their frequency in the descriptions of similar images and their generality in the overall dataset. Experimental result by human evaluation shows that our image descriptions include fewer errors and visually relevant describe images. However, experimental result by automatic evaluation indicates that our image descriptions tend to be too concise. Because our approach uses similar image retrieval to select the nouns as objects and their attribute phrases, our generated image descriptions are also affected by the quality of image retrieval system and the variations of dataset of images and descriptions, houever our approach outputs the visually relevant descriptions in the noisy dataset.

6. REFERENCES

[1] N. Chen, Q.Y. Zhou, and V. Prasanna. Understanding web images by object relation network. In Proceedings of the 21st international conference on World Wide Web, pages 291–300. ACM, 2012.

[2] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. Computer Vision–ECCV 2010, pages 15–29, 2010.

[3] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, and T.L. Berg. Baby talk: Understanding and

generating simple image descriptions. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 1601–1608. IEEE, 2011.

[4] P. Kuznetsova, V. Ordonez, A.C. Berg, T.L. Berg, and Y. Choi. Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, pages 359–368, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[5] V. Ordonez, G. Kulkarni, and T.L. Berg. Im2text: Describing images using 1 million captioned photographs. In Proceedings of NIPS, 2011.

[6] B.Z. Yao, X. Yang, L. Lin, M.W. Lee, and S.C. Zhu. I2t: Image parsing to text description. Proceedings of the IEEE, 98(8):1485–1508, 2010.