Image Collector III: A Web Image-Gathering System with Bag-of-Keypoints

Keiji Yanai
The University of Electro-Communications
Chofu, Tokyo, 182-8585 JAPAN
yanai@cs.uec.ac.jp

ABSTRACT

We propose a new system to mine visual knowledge on the Web. There are huge image data as well as text data on the Web. However, mining image data from the Web is paid less attention than mining text data, since treating semantics of images are much more difficult. In this paper, we propose introducing a latest image recognition technique, which is the bag-of-keypoints representation[1], into Web image-gathering task. By the experiments we show the proposed system outperforms our previous systems and Google Image search greatly.

Categories and Subject Descriptors: I.4 [Image Processing and Computer Vision]: Miscellaneous

General Terms: Algorithms, Experimentation, Measurement

Keywords: Web image mining, image recognition, bag-of-keypoints

1. INTRODUCTION

Because of the recent growth of the World Wide Web, we can easily gather substantive quantities of image data. Our goal is to mine such data for visual content. In particular, we wish to build a large scale data set consisting of many highly relevant images for each of thousands of concepts. To realize that, we have proposed several Web image gathering systems so far [3, 4, 5]. But they employed relatively simple image recognition technique.

In this paper we present a new system employing an up-to-date image recognition technique for visual object categorization / recognition, which is bag-of-keypoints representation [1]. The bag-of-keypoints representation got popular recently in the research community of computer vision. It is proven that it has excellent ability to represent image concepts in the context of visual object categorization/recognition in spite of its simplicity [1].

The basic idea of the bag-of-keypoints representation is that a set of local image patches is sampled by an interest point detector or randomly, and a vector of visual descriptors is evaluated by Scale Invariant Feature Transform (SIFT) descriptor [2] on each patch. The resulting distribution of description vectors is then quantified by vector quantization against a pre-specified codebook, and the quantified distribution vector is used as a characterization of the image. As a classifier to classify images associated with quantified vectors as relevant or irrelevant, we use an SVM classifier.

In this paper, we propose an Web image-gathering system with the bag-of-keypoints model. By the experiments we show the new system outperforms our previous systems greatly.

2. OVERVIEW OF PROPOSED SYSTEM

The proposed system gathers images associated with the keywords provided by a user fully automatically. Therefore, an input of the system is just keywords, and the output is several hundreds or thousands images associated with the keywords.

Our proposed system consists of two stages, which are a collection stage and a selection stage. In this paper, we modify only the selection stage of our previous system[4].

In the collection stage, we gather many images and HTML documents related to the given keywords using Web search engines. We perform evaluation of the relevancy of images by analyzing associated HTML documents. According to the relevancy of images to the given keywords, we divide images into two groups: images in group A are highly relevant to the keywords, and others are classified into group B. The possibility that images in group A are relevant is high, so that we use them as training data of a SVM classifier in the next stage, although they includes a small number of irrelevant ones. The detail is described in [4, 5].

In the selection stage, we select relevant images from all the downloaded images by employing image analysis. In this paper, we use the bag-of-keypoints model [1] as an image representation and an SVM classifier as a classification method. In general, to use machine learning methods like an SVM to select true images, we need labeled training images. However, we do not want to pick up good images by hand. Instead, we regard images classified into group A as training images, although they always include some irrelevant images. In this paper, we provide a classifier with all group-A images as relevant training images.

In the selection stage, first we convert all the downloaded images into feature vectors based on the bag-of-keypoints representation, and then train an SVM classifier with all the vectors in the group A as training data. Next, we classify all the vectors in the group A and B as relevant or irrelevant with the trained SVM. Finally, we can get only images classified as relevant to the provided keywords as a result. The detail of this processing is as follows:

1. Sample many image patches from each image
2. Extract patch feature vectors from all the points by SIFT descriptor [2]
3. Generate codebooks with k-means clustering over extracted patch feature vectors
4. Assign all patch feature vectors to the nearest codebooks, and convert a set of patch feature vectors for each image into one histogram vector of assigned codebooks.
5. Train an SVM classifier with all the histogram vectors in the group A as training data.
6. Classify all the histogram vectors of downloaded images as relevant or irrelevant with applying the trained SVM.

The main idea of the bag-of-keypoints model is representing images as collections of independent local patches, and vector-quantizing them as histogram vectors [1].

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

Table 1: Results by the CBIR-based method [4], the GMM-based probabilistic method [5] and the proposed system. This table describes the precision of the 500 output images of Google Image Search which are ranked from 1 to 500, the number of raw images collected from the Web, the number of selected images out of them by the two old methods. Numerical values in () represent the precision and the recall.

	Goo.	raw images			CBIR [4]	region-based [5]	bag-of-keypoints (proposed system)		
concepts	prec.	A	B	A+B	A+B	A+B	A	B	A+B
sunset	79.8	790 (67)	710 (44)	1500 (55.3)	828 (62.2, 62.1)	**636 (91.0, 70.2)**	441 (94, 78)	113 (92, 34)	**564 (93.3, 62.5)**
mountain	48.8	1950 (88)	3887 (71)	5837 (79.2)	3423 (82.6, 61.2)	**3510 (89.0, 65.0)**	1628 (94, 89)	1133 (92, 46)	**2761 (93.7, 68.7)**
Chinese noodle	65.2	901 (78)	1695 (55)	2596 (66.6)	1492 (71.0, 61.3)	**1266 (77.0, 53.2)**	572 (84, 68)	448 (94, 33)	**1020 (86.9, 54.2)**
waterfall	72.4	2065 (71)	2584 (70)	4649 (70.3)	3281 (71.4, 71.7)	**3504 (76.8, 74.6)**	1728 (80, 94)	1535 (86, 62)	**3263 (82.3, 82.0)**
beach	63.2	768 (69)	1155 (62)	1923 (65.5)	1128 (67.3, 60.3)	**983 (73.3, 62.5)**	440 (84, 70)	262 (93, 24)	**702 (86.5, 48.7)**
flower	65.6	576 (72)	1418 (67)	1994 (69.6)	952 (79.3, 54.4)	**758 (71.9, 41.0)**	360 (84, 73)	348 (94, 18)	**708 (86.7, 45.0)**
lion	44.0	511 (87)	1548 (49)	2059 (66.0)	967 (71.0, 50.5)	**711 (69.4, 53.6)**	414 (87, 81)	375 (73, 18)	**789 (85.5, 56.1)**
apple	47.6	1141 (78)	2137 (59)	3278 (64.3)	1495 (68.8, 48.8)	**1252 (67.2, 37.7)**	759 (85, 73)	212 (84, 20)	**971 (85.3, 38.5)**
baby	39.4	1833 (56)	1738 (53)	3571 (54.5)	1831 (55.1, 51.8)	**1338 (63.9, 45.9)**	1441 (54, 76)	601 (61, 29)	**2042 (55.7, 58.4)**
notebook PC	60.2	781 (57)	1756 (32)	2537 (43.6)	1290 (46.9, 54.6)	**867 (56.0, 47.6)**	612 (58, 80)	602 (45, 42)	**1214 (55.0, 66.3)**
TOTAL/AVG.	58.6	11316 (72)	18628 (56)	29944 (62.2)	16687 (66.0, 57.7)	**14825 (73.5, 55.1)**	7926 (80, 78)	5371 (81, 32)	**13297 (81.1, 58.0)**

3. EXPERIMENTAL RESULTS

We made experiments for the following ten concepts independently: beach, sunset, flower, waterfall, mountain, lion, apple, baby, note-PC, and Chinese noodle. For only "lion" and "apple", actually we added subsidiary keywords "animal" and "fruit" to restrict its meaning to "lion of animal" and "apple of fruit" in the collection stage, respectively.

In the collection stage, we gathered around 5000 URLs for each concept from both Google Search and Yahoo Web Search which are not "image search" but "text search". The exact numbers vary depending on concepts, since we excluded duplicate URLs from the URL list for each category.

Table 1 shows the results of the collection stage, namely raw images, and we added to it the evaluation of the results of Google Image Search and our previous system which employs the CBIR-based image selection method [4] and GMM-based probabilistic method [5] for comparison. The results of the collection stage consists of the number of images downloaded from the Web with only HTML analysis and their precision. To compute the precision and the recall, we randomly selected 500 images from the images of each concepts and checked their relevancy by the subjective evaluation. Note that for the downloaded images we cannot estimate the recall, since the denominator to estimate it corresponds to the number of images associated to the given concept on whole the Web and we cannot get to know it. Regarding the results of Google Image Search, we show the precision of output images ranked between 1 and 500 in the table. The average precision of raw images, 62.2%, was slightly superior to the average precision of top 500 results of Google images, 58.6%, while we collected about 3000 images a concept. This shows that our original image collection method is better than Google Image Search.

Table 1 also shows the number, the precision and the recall of the results by the proposed method by the bag-of-keypoints model and SVM. In the experiments, we used the parameter setting so that the recall rates are close to the recall rate by two old methods shown in Table 1 for easy comparison. Note that in the Web image gathering task, the recall rate is less important than the precision rate, since the more Web sites we crawl, the more images we can get easily. So we mainly evaluate the system performance by the precision below.

In case of (1), we obtained the 81.1% precision on the average, which outperformed the 66.0% precision by the CBIR method and the 73.5% precision by the GMM-based probabilistic method. Except "baby" and "notebook PC", the precision of each concept were also improved. Especially, in case of "flower", "lion" and "apple", the precision were improved prominently. This shows that the bag-of-keypoints representation is very effective to classify "object" images. On the other hand, the precisions of "baby" and "notebook PC" were not good, which were less than the precision by the probabilistic methods. This is because we used all the A-group images as positive training samples, and for these two concept the precision of the raw A-group images were 56% and 57%, respectively. In short, training data for two concepts contained too many irrelevant samples. That is why the precisions were not improved. To overcome that, we need to prepare better raw group-A images or to develop a mechanism to remove irrelevant training samples.

We have prepared the Web site to show the experimental results we provided in this paper. The URL is as follows:
http://mm.cs.uec.ac.jp/yanai/www07/

4. CONCLUSIONS

In this paper, we described a new system employing the bag-of-keypoints representation [1] which was paid much attention to as a new excellent image representation for visual object categorization / recognition. As a classifier, we use an SVM classifier. In the experiments for ten concept keywords, we obtained the 82.2% precision on the average in case of 1000 codebooks with random sampling, which outperformed the 66.0% precision by the CBIR method and the 73.5% precision by the GMM-based probabilistic method. The experimental results shows that the bag-of-keypoints representation is very effective to classify "object" images as well as "scene" images. However, when the precision of the training data is not enough, our methods cannot improve the precision since we use an SVM classifier.

As future work, we plan to prepare better raw group-A images by improving HTML analysis methods and combining query keywords for Web search engines with effective subsidiary keywords, and we need to study how to remove irrelevant data in training data or how to learn from imperfect training image data gathered from the Web.

5. REFERENCES

[1] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Proc. of ECCV Workshop on Statistical Learning in Computer Vision*, pages 1–22, 2004.

[2] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[3] K. Yanai. Generic image classifi cation using visual knowledge on the web. In *Proc. of ACM International Conference Multimedia*, pages 67–76, 2003.

[4] K. Yanai. Image collector II : An over-one-thousand-image-gathering system. In *Proc. of the Twelfth International World Wide Web Conference*, 2003.

[5] K. Yanai and K. Barnard. Probabilistic web image gathering. In *Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 57–64, 2005.