WEB IMAGE GATHERING WITH REGION-BASED BAG-OF-FEATURES AND MULTIPLE INSTANCE LEARNING

Keiji Yanai

Department of Computer Science, The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585 JAPAN

ABSTRACT

We propose a new Web image gathering system which employs the region-based bag-of-features representation and multiple instance learning. The contribution of this work is introducing the region-based bag-of-features representation into an Web image gathering task where training data is incomplete and having proved its effectiveness by comparing the proposed method with the normal whole-image-based bag-of-features representation.

In our method, first, we perform region segmentation for an image, and next we generate a bag-of-features vector for each region. One image is represented by a set of bag-of-features vectors in this paper, while one image is represented by just one bag-of-features vector in the normal bag-of-features representation which is very popular for visual object categorization tasks recently.

Several works on Web image selection with bag-of-features have been proposed so far. However, in case that the training data includes much noise, sufficient results could not be obtained. In this paper, we divide images into regions and classify each region with multiple-instance support vector machine (mi-SVM) instead of classifying whole images. By this region-based classification, we can separate foreground regions from background regions and achieve more effective image training from incomplete training data. By the experiments, we show that the results by the proposed methods outperformed the results by the whole-image-based bag-of-visual-words and the normal support vector machine.

1. INTRODUCTION

The recent explosive growth of the World Wide Web has enabled easy acquisition of a huge number of various kinds of images. Our goal is to build a large scale data set consisting of many highly relevant images for each of thousands of concepts by gathering images from the Web. To realize that, we have proposed several Web image gathering systems so far [1, 2, 3], and similar works have been proposed in the literature [4, 5, 6].

Most of object recognition work assumes that complete training images are available. On the other hand, since Web image gathering system selects training images automatically based on HTML analysis, complete training images cannot be available and training data always include some noise. Learning from incomplete training data is the biggest difference from common image recognition task, and it causes difficulty of this kind of work.

To prepare training images automatically without human intervention, we proposed the HTML-based automatic positive training image selection based on simple heuristics on HTML tags [1]. Although improvement of the HTML-based method to select training images is possible, in this paper we focus on visual learning from incomplete training data.

In our previous work [3], we proposed introducing the bag-of-features representation and a support vector machine into Web image gathering task. The proposed method worked very well in case that the precision of training data is relatively high, but it did not worked well in case that the precision of training data is less than about 60%. To overcome such the problem, we propose introducing region-based bag-of-features into the Web image gathering task in this paper. This is inspired by the work by Ravinovich et al.[7]. They used the bag-of-features representation [8] to represent each segmented region for multiple object recognition task where one image contains several kinds of objects such as car, building, road and sky at the same time. Our objective is different from theirs, since we like to examine if a downloaded Web image is relevant to the given concept without human intervention, while their work is related to object recognition tasks where perfect training data is available. Our intention of using region segmentation is to enable us to learn from noisy data set by separating foregrounds from backgrounds. In this paper, for background separation, we use multiple-instance support vector machine (mi-SVM) [9].

Actually, in our other work [2], we proposed the method with region segmentation and a probabilistic model based on a Gaussian mixture model, which made it possible to separate foreground regions from background regions. Regions are represented by their color, shape and texture. The method proposed in this paper can be regarded as the extension of the work [2].

As a similar work, Schroff et al. examined effectiveness of combination of bag-of-features [8] and SVM extensively [6]. Their works proved that SVM with soft margin worked well even under the condition where training data contains noise. However, they used not region segmentation but the normal bag-of-features as whole image representation.

In this paper, we propose a new Web image gathering system which employs the region-based bag-of-features representation and multiple-instance support vector machine (mi-SVM). The contribution of this work is introducing the region-based bag-of-features representation and multiple-instance SVM into an Web image gathering task where training data is incomplete. We use our original HTML-based automatic selection methods as a method to construct initial pseudo-training data, and we use the mi-SVM [9] as a classifier. By the experiments, we show its effectiveness by comparing it with the normal whole-image-based bag-of-features representation and the normal SVM.

Note that the objective of our image gathering is absolutely different from ones of the other Web image search systems including commercial Web image search engines. While their objective is searching for highly relevant but relatively a small number of images, ours is to gather a large number of relevant images to build image concept database. So that we call our system not Web image search system but "Web image gathering system".

The rest of this paper is organized as follows: In Section 2 we overview our system which consists of a collection stage and a selection stage. In Section 3 we explain the detail of the selection stage which performs image-based image selection. In Section 4 we presents the experimental results and evaluations, and in Section 5 we conclude this paper.

2. OVERVIEW OF WEB IMAGE SELECTION

We assume that the method we propose in this paper is used in the image selection stage of the Web image-gathering system [1, 2, 3]. The system gathers images associated with the keywords given 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. The system consists of two stages: the collection stage and the selection stage.

In the collection stage, the system carries out HTML-text-based image selection which is based on the method we proposed in [1]. The basic idea on this stage is to gather as many images related to the given keywords as possible from the Web with Web text search engines such as Google and Yahoo, and to select candidate images which are likely to be associated with the given keywords by analysis of surrounding HTML text based on simple heuristics. Particularly high-scored images among the candidate images are selected as pseudo-training images for training the SVM. To explain simple HTML analysis briefly, if either ALT tags, HREF link words or image file names include the given keywords, the image is regarded as a pseudo-training image. If the other tags or text words which surround an image link include the given keywords, the image is regarded as a normal candidate image. Although the former rule to select training images is strongly restrictive, this simple rule can find out highly relevant images which can be used as pseudo-training samples by examining a great many image gathered from the Web. The detail on the collection stage is described in [1].

In the selection stage, we perform an image-feature-based image selection using the region-based bag-of-features representation as an image representation and an multiple-instance SVM classifier [9] as a classification method. Since mi-SVM is a supervised machine learning method, we use pseudo-training images selected automatically in the collection stage as positive training samples. As negative training samples, we use random images gathered from the Web by sending random keywords to the Web image search engine. Note that all pseudo-training images are also part of candidate images at the same time, since pseudo-training images are also Web images and contain several irrelevant images which should be removed.

3. SELECTION STAGE

In this section, we describe the detail of the selection stage, especially about region-based bag-of-features and multiple-instance SVM.

The main idea of the bag-of-features representation [8] is representing images as collections of independent local patches, and vector-quantizing them as histogram vectors. Before constructing the bag-of-features vector, we apply region segmentation for all the images. As a region segmentation method, we use JSEG [10] after adjusting the parameters so as to generate about eight regions per image on average.

The main steps of the method are as follows:

1. Carry out region segmentation with the JSEG algorithm.
2. Sample 3000 patches per image randomly in the same way as [11].
3. Generate feature vectors for the sampled patches by the SIFT descriptor [12].
4. Construct a codebook with k-means clustering over all the extracted feature vectors. A codebook is constructed for each concept independently. We set k as 1000.
5. Assign all SIFT vectors to the nearest codeword of the codebook, and convert a set of SIFT vectors for each region into one k-bin histogram vector regarding assigned codewords. In addition, background images which are prepared as negative training samples in advance are also divided into regions and converted the sets of SIFT vectors extracted from regions into k-bin histograms based on the same codebook.
6. Classify candidate images with the mi-SVM [9], which is carried out by repeating training and selection of positive samples with the normal SVM. The detail is described in Section 3.1.

3.1. Classifying Regions with Multiple-Instance SVM

To classify each region of candidate images as relevant and irrelevant, we use a multiple instance SVM (mi-SVM) classifier [9]. The mi-SVM is a support vector machine modified for multiple instance setting.

Under the multiple instance setting, training class labels are associated with a set of instances instead of individual instances. A positive set, which is called as a "positive bag", has one positive instance at least, while a negative set, which is called as a "negative bag", has only negative instances. This multiple instance setting fits well with the situation where an image consists of several foreground regions and background regions. Since we can regard foreground and background regions as positive and negative instances, respectively, by using multiple-instance learning methods we can classify regions into foregrounds and backgrounds.

In this paper, we use multiple instance SVM (mi-SVM) classifier [9] as a classifier for the multiple instance setting. The mi-SVM is carried out by iterating a training step and a classification step twice using the normal SVM.

At the first iteration, we prepare pseudo-positive training images and random negative images as positive bags and negative bags, respectively. Next, we train the normal SVM treating with all the regions of positive bags as positive training samples and all the regions of negative bags as negative

training samples. After that, we apply the trained SVM to all the vectors of all the regions of positive bags and classify the regions of positive bags into positive regions or negative regions.

At the second iteration, we use only regions classified into positive ones in the first step as positive samples and all the other regions as negative samples, and train the SVM again. Finally we apply the trained SVM to all the regions of candidate images, and obtain the output value of the SVM for each region which corresponds to the distance between the given vector and the discriminative hyper-plane in the context of SVM.

For each candidate image, we sort regions in the descending order of the output value of the SVM with the image, and select the top several regions so that the total relative size of selected regions gets more than the predefined threshold α. In the experiments, we set 0.3 to α. Finally, we obtain the final score of the image as the weighted average of the SVM output value of the selected regions. Weighting is carried out in the proportion to the size of the regions. Note that weighting based on region size is our own modification of the original mi-SVM, which selects the maximum score of regions as the final score of a bag regardless of the size of regions. By region-size-based weighting, we prevent the score of tiny regions from reflecting the final output greatly.

4. EXPERIMENTAL RESULTS

We made experiments for the following fifteen concepts independently: sunset, mountain, waterfall, beach, flower, lion, apple, baby, notebook-PC, Chinese noodle, airplane, guitar, leopard, motorbike and watch. The first four concepts are "scene" concepts, and the rest are "object" concepts. Since the last five concepts are commonly used in several works such as [5] and [6], we used them for comparison. 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 obtained around 5000 URLs for each concept from several Web search engines including Google Search and Yahoo Web Search. The exact numbers vary depending on concepts, since we excluded duplicate URLs from the URL list for each concept.

Table 1 shows the precision of top 100 output images of Google Image Search for comparison, the number and the precision of pseudo-training images and candidate images, and the results of image selection by the region-based probabilistic method employing a Gaussian Mixture Model (GMM) [2], by whole-image-based bag-of-features (normal BoF) with SVM [3] and by region-based bag-of-features (region BoF) with mi-SVM. Note that in [2] we used ten concepts, so that we show the results for the same ten concepts in this table. The GMM-based method employs region segmentation in the similar way to the proposed method, but in the GMM-based method an image are represented as a collection of region features regarding color, texture and shape of regions. In the experiments, all the precision of the results except for positive training and candidate images are evaluated at 15% recall. This evaluation standard is the same as in [6]. Table 2 shows the precision (at 15% racall) of the results by the whole-image-based bag-of-features (normal BoF) representation and by the region-based bag-of-features (region BoF) for the additional five concepts. In the bottom of the row of Table 2, the average of the precision over 15 concepts is shown for the results by normal BoF and region-based BoF.

Regarding the results of Google Image Search, we show the precision of output images ranked between 1 and 100 in the table. The average precision of candidate images, 62.2%, was almost equivalent to the average precision of the top 100 results of Google images, 63.6%, while we collected about 3000 images a concept. This shows that the simple heuristic method employed in the collection stage worked well.

In case of "region-based BoF", we obtained the 90.7% precision on the average of the ten concepts and the 92.8% precision on the average of the fifteen concepts, which outperformed the 80.0% precision by the GMM-based probabilistic method and the 80.6% precision by "normal BoF". Although for most of the concepts the precision of "region-based BoF" were about 80% to 90%, the precision still remained low for only "baby". This is because "baby" images include many "baby" related images such as baby toys, baby furniture and animal babies. When evaluating, we regarded them as negative "baby" images, so that the precision remained low. If we regard them as "baby" images, the precision rate will be 94.6% and 89.6% for "normal BoF" and "region-based BoF", respectively.

For the additional five concepts and "notebook PC", the results were improved greatly compared to "normal BoF", because all of them are "object" concepts, and classification of foreground and background regions separately worked effectively. On the other hand, for scene concepts such as "mountain" and "waterfall", the results by the normal BoF were better than the results by the region-based BoF, since the combination of both foregrounds and backgrounds is important to represent "scene" concepts. For scene concepts, background separation is not likely to be needed.

We show parts of the result images of "notebook PC" in Figure 1, "airplane" in Figure 2, respectively. With mi-SVM, we can get to know which regions are likely to be foreground regions. In these figures, labels are added to the regions judged as foreground regions.

Table 2 also shows the results of Fergus et al.[5] and Schroff et al.[6]. All of our results by "region-based BoF" for the five concepts outperformed them greatly. Especially, the precision on "airplane" shown in Figure 2 were improved greatly, since our method tries to classify images by focusing only foreground regions, while they used the whole-image bag-of-features representation. This demonstrates the effectiveness of our proposed method.

Unfortunately we can show only small parts of the result images due to space limitation. Instead we have prepared the Web site to show the experimental results we provided in this paper. The URL is as follows:

http://img.cs.uec.ac.jp/yanai/icme09/

5. CONCLUSIONS

In this paper, we described a new system employing the region-based bag-of-features representation and the multi-

Table 1. The precision of top 100 output images of Google Image Search for comparison, the number and the precision (at 15% recall) of pseudo-training images and candidate images which are selected automatically in the collection stage, and the results of image selection by the region-based probabilistic method employing GMM [2], by whole-image-based bag-of-features (normal BoF) with SVM [3] and by region-based bag-of-features (region BoF) with mi-SVM. All the results are represented in the precision at 15% recall.

concepts	Google	positive	candidate	GMM [2]	normal BoF [3]	region-based BoF [this paper]
sunset	85	790 (67)	1500 (55.3)	100.0	98.0	100.0
mountain	57	1950 (88)	5837 (79.2)	96.5	100.0	92.0
waterfall	78	2065 (71)	4649 (70.3)	82.0	96.7	85.9
beach	67	768 (69)	1923 (65.5)	75.0	99.0	99.0
flower	71	576 (72)	1994 (69.6)	78.5	93.8	91.2
lion	52	511 (87)	2059 (66.0)	74.6	84.6	94.3
apple	49	1141 (78)	3278 (64.3)	81.0	93.1	95.7
baby	39	1833 (56)	3571 (54.5)	70.7	66.0	59.3
notebook PC	70	781 (57)	2537 (43.6)	70.5	53.7	91.7
Chinese noodle	68	901 (78)	2596 (66.6)	70.9	96.8	98.4
TOTAL/Average	**63.6**	**11316 (72)**	**29944 (62.2)**	**80.0**	**88.2**	**90.7**

Table 2. The precision (at 15% recall) of the results by whole-image-based bag-of-features (normal BoF) with SVM [3], by region-based bag-of-features (region BoF) with mi-SVM, by Fergus et al. [5] and by Schroff et al. [6].

concepts	normal BoF [3]	region-based BoF [this paper]	Fergus [5]	Schroff [6]
airplane	58.9	100.0	57	45
guitar	63.2	90.0	50	72
leopard	46.7	100.0	59	72
motorbike	67.5	95.9	71	81
watch	91.4	98.0	88	97
Average (5 concepts)	**65.5**	**96.8**	65.0	73.4
Average (15 concepts†)	**80.6**	**92.8**		

†...includes the ten concepts shown in Table 1

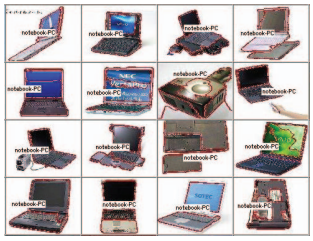

Fig. 1. "Notebook PC" images.

Fig. 2. "Airplane" images.

ple instance SVM. The combination of region-based bag-of-features and the mi-SVM classifier improved the result compared with normal bag-of-features. In the experiments for ten concept keywords, we obtained the 92.8% precision on the average. which outperformed the 80.6% precision by the normal bag-of-features. Especially, the results of "notebook PC" were improved greatly, although the precision of its initial pseudo-positive samples are less than 60%.

As future work, we plan to prepare better initial training images by improving HTML analysis methods and combining query keywords for Web search engines with effective subsidiary keywords, and plan to investigate how to remove irrelevant data in training data or how to learn from imperfect training data more effectively.

6. REFERENCES

[1] K. Yanai, "Generic image classification using visual knowledge on the web," in *Proc. of ACM International Conference Multimedia*, 2003, pp. 67–76.

[2] K. Yanai and K. Barnard, "Probabilistic Web image gathering," in *Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2005, pp. 57–64.

[3] K. Yanai, "Image collector III: A web image-gathering system with bag-of-keypoints," in *Proc. of the International World Wide Web Conference*, 2007.

[4] R. Fergus, P. Perona, and A. Zisserman, "A visual category filter for google images," in *Proc. of European Conference on Computer Vision*, 2004, pp. 242–255.

[5] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *Proc. of IEEE International Conference on Computer Vision*, 2005, pp. 1816–1823.

[6] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," in *Proc. of IEEE International Conference on Computer Vision*, 2007.

[7] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proc. of IEEE International Conference on Computer Vision*, 2007.

[8] 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*, 2004, pp. 59–74.

[9] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support Vector Machines for Multiple-Instance Learning," in *Advances in Neural Information Processing Systems*, 2003, pp. 577–584.

[10] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 800–810, 2001.

[11] E. Nowak, F. Jurie, W. Triggs, and M. Vision, "Sampling strategies for bag-of-features image classification," in *Proc. of European Conference on Computer Vision*, 2006, pp. IV:490–503.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.