Automatic Web Image Selection with a Probabilistic Latent Topic Model

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

ABSTRACT

We propose a new method to select relevant images to the given keywords from images gathered from the Web based on the Probabilistic Latent Semantic Analysis (PLSA) model which is a probabilistic latent topic model originally proposed for text document analysis. The experimental results show that the results by the proposed method are almost equivalent to or outperform the results by existing methods. In addition, it is proved that our method can select more various images compared to the existing SVM-based methods.

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

General Terms: Algorithms, Experimentation

Keywords: Web image mining, image recognition

1. INTRODUCTION

Because of the recent growth of the World Wide Web, we can easily gather huge amount of image data. However, raw outputs of Web image search engines contain many irrelevant images, since they do not employ image analysis and basically rely on only HTML text analysis to rank images. Our goal is to gather large amount of relevant images to given words. In particular, we wish to build a large scale generic image database consisting of many highly relevant images for each of thousands of concepts, which can be used as huge ground truth data for generic object recognition research. To realize that, we have proposed several Web image gathering systems employing image recognition methods so far [5, 6, 7].

In this paper, we apply Probabilistic Latent Semantic Analysis (PLSA) to Web image gathering task. Recently, PLSA is applied to object recognition task as a probabilistic generative model [4]. However, PLSA is not applied to Web images except [1]. The difference between this paper and [1] is that in [1] they selects just one topic as a relevant topic while our proposed method selects relevant images based on the mixture of positive topics. This can be regarded as an extension of our previous work [6], which employed region segmentation and a probabilistic model based on a Gaussian mixture model (GMM). In [6], an image is represented as a set of region feature vectors such as color, texture and shape, while in this paper we use bag-of-visual-words representation [2] to represent an image. A method to recognize images based on the mixture of topics has already proposed in [4]. Our work can be regarded as the Web image version of that work.

In this paper, we propose a fully automated PLSA-based Web image selection method for the Web image-gathering task. The method employs the bag-of-visual-words as image representation and a PLSA-based topic mixture model as a probabilistic model. Our main objective is to examine if the bag-of-visual-words model and the PLSA-based model are also effective for the Web image gathering task where training images always contains some noise.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. OVERVIEW OF THE METHOD

We assume that the method we propose in this paper is used in the image selection stage of the Web image-gathering system [6, 7]. 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 before [5]. 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 probabilistic model. 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 [5].

In the selection stage, the proposed model is trained with the pseudo-training images selected automatically in the collection stage, and is applied to select relevant images from the candidate images. 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.

As an image representation, we adopt the bag-of-visual-words representation [2]. It has been proved that it has the excellent ability to represent image concepts in the context of visual object recognition in spite of its simplicity. The basic idea of the bag-of-visual-words representation is that a set of local image patches is sampled by an interest point detector or a grid, and a vector of visual descriptors is evaluated by Scale Invariant Feature Transform (SIFT) descriptor [3] on each patch. The resulting distribution of

Table 1: The precision of top 100 output images of Google Image Search, the number and the precision (at 15% recall) of positive images and candidate images which are selected automatically in the collection stage, the results of image selection by the region-based probabilistic method employing GMM [6] and the bag-of-visual-words-based method employing SVM [7] for comparison and results by the proposed PLSA-based methods with five different k. k is the number of topics.

concepts	Google result	positive images	candidate images	GMM	SVM	PLSA(proposed method) k=10	k=20	k=30	k=50	k=100	BEST
sunset	85	790 (67)	1500 (55.3)	100.0	98.0	95.1	96.0	96.0	95.1	**97.0**	**97.0**
mountain	57	1950 (88)	5837 (79.2)	96.5	100.0	93.9	**96.5**	**96.5**	**96.5**	**96.5**	**96.5**
waterfall	78	2065 (71)	4649 (70.3)	82.0	90.7	75.3	**78.1**	75.3	76.8	74.5	**78.1**
beach	67	768 (69)	1923 (65.5)	75.0	99.0	92.5	94.2	**96.1**	94.2	93.3	**96.1**
flower	71	576 (72)	1994 (69.6)	78.5	91.9	**83.9**	82.3	80.8	81.3	81.3	**83.9**
lion	52	511 (87)	2059 (66.0)	74.6	85.7	82.5	66.7	64.7	84.6	**85.7**	**85.7**
apple	49	1141 (78)	3278 (64.3)	81.0	90.7	**88.2**	82.7	84.8	87.0	83.8	**88.2**
Chinese noodle	68	901 (78)	2596 (66.6)	70.9	95.3	93.8	90.9	89.5	**95.2**	**95.2**	**95.2**
TOTAL/AVG.	65.9	8702 (76)	23836 (66.5)	82.4	93.9	88.2	85.9	85.5	**88.8**	88.4	**90.1**

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.

The proposed model is based on Probabilistic Latent Semantic Analysis (PLSA). PLSA is originally an unsupervised latent topic model. First, we apply the PLSA method to the candidate images with the given number of topics, and get the probability of each topic over each image, $P(z|I)$. Next, we calculate the probability of being positive or negative regarding each topic, $P(pos|z)$ and $P(neg|z)$ using pseudo-training images, assuming that all other candidates images than pseudo positive images are negative samples. Here, "positive topic" means that the latent topic generates images relevant to the given keywords, and "negative topic" means that the latent topic generates irrelevant images. Finally, the probability of being positive over each candidate image, $P(pos|I)$, is calculated by marginalization over topics:

$$P(pos|I) = \sum_{z \in Z} P(pos|z)P(z|I) \quad (1)$$

, where $z \in Z$ represents latent topics, the number of which is decided by the given number k. We can rank all the candidate images based on this probability, $P(pos|I)$, and obtain the final result.

3. EXPERIMENTAL RESULTS

We made experiments for the following eight concepts independently: sunset, mountain, waterfall, beach, flower, lion, apple and Chinese noodle. The first four concepts are "scene" concepts, and the rest are "object" concepts.

In the collection stage, we obtained around 5000 URLs for each concept from several Web search engines including Google Search and Yahoo Web Search.

Table 1 shows the precision of top 100 output images of Google Image Search for comparison, the number and the precision of positive images and candidate images, and the results of image selection by the region-based probabilistic method employing GMM [6] and the bag-of-visual-words-based method employing SVM [7] for comparison. In the experiments, all the precision of the results except for positive and candidate images are evaluated at 15% recall.

The 7th to 11th column of Table 1 shows the results of the precision of the PLSA-based image selection when the number of topics k varied from 10 to 100. In terms of the best results, the precision of each keyword is almost equivalent to the precision by SVM and outperforms GMM and Google Image Search. As shown in Table 1, the average of the precision of positive images is 76%, while the average of the precision of candidate images is 65%. Although their difference is about 10% and it is not so large, our proposed strategy to estimate positive and negative topics worked well in the most case.

Regarding the number of topics k when the best result was obtained, there is not a prominent tendency. For future work, we need to study how to decide the number of topics, which sometimes influence the result greatly. For example, in case of "apple", the precision was 85.7% for $k = 100$, while the precision was 64.7% for $k = 30$.

The biggest difference to [7] is that our higher-rank results include various images as shown in Fig.1, while ones by SVM [7] include similar and uniform images as shown in Fig.2. This is because our proposed method is based on the mixture of the topics.

Figure 1: "Mountain" by PLSA.

Figure 2: "Mountain" by SVM.

We have prepared the experimental results on the Web: `http://mm.cs.uec.ac.jp/yanai/www08/`

4. REFERENCES

[1] 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*, pages 1816–1823, 2005.

[2] 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 59–74, 2004.

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

[4] F. Monay and D. Gatica-Perez. Modeling semantic aspects for cross-media image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1802–1817, 2007.

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

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

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