Region-Based Automatic Web Image Selection

Keiji Yanai

Department of Computer Science,
The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi,
Tokyo, 182-8585 JAPAN
yanai@cs.uec.ac.jp

Kobus Barnard

Computer Science Department,
University of Arizona
Tucson, AZ, 85721 USA
kobus@cs.arizona.edu

ABSTRACT

We propose a new Web image selection method which employs the region-based bag-of-features representation. The contribution of this work is (1) to introduce the region-based bag-of-features representation into an Web image selection task where training data is incomplete, and (2) to prove its effectiveness by experiments with both generative and discriminative machine learning methods. In the experiments, we used a multiple-instance learning SVM and a standard SVM as discriminative methods, and pLSA and LDA mixture models as probabilistic generative methods.

Several works on Web image filtering task 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 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-features.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Miscellaneous

General Terms

Algorithm, Experimentation

Keywords

Web image mining, region-based, multiple instance learning, pLSA, LDA, SVM

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 contains 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 the same end, several works on image gathering from the Web to build an image database has been proposed so far [30, 9, 11, 10, 33, 15, 21, 32, 26, 29]. Most of these works employed object categorization / recognition methods to select relevant images to given keywords from "raw" images collected from the Web using Web image search engines.

In general, object categorization / recognition methods assume that noise-free training data is available, since they usually employs supervised machine learning methods such as SVM and boosting. In the Web image search / gathering task, a fully automated processing is desirable to archive the goal that is to build a large-scale image database. Therefore, it is needed to prepare training images automatically without human intervention under the condition that only text keywords are given in advance. In the fully automatic Web image gathering task, since a system selects training images 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 to common image recognition task, and it causes difficulty of this kind of tasks.

To learn from incomplete data on the Web automatically, several methods have been proposed so far [10, 15, 21, 32, 26]. They employed the bag-of-features (BoF) representation [6] as image representation, which has made a great success on object recognition. In these works, bag-of-features vectors are constructed from whole images, so that they sometimes does not work well in case that regions of objects are relatively small and backgrounds are large in images. 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.[20]. They used the bag-of-features representation 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 or not. Our intention of using region segmentation is to enable us to learn from noisy data set by separating foregrounds from backgrounds.

MIR'10, March 29–31, 2010, Philadelphia, Pennsylvania, USA.

In this paper, we propose a new Web image gathering method which employs the region-based bag-of-features representation. The novelty of this work is to introduce the region-based bag-of-features representation into an Web image gathering task where training data is incomplete. To show the effectiveness of the region-based bag-of-features in the noisy Web images, we made the experiments using a multiple-instance learning SVM and a standard SVM as discriminative methods, and pLSA and LDA mixture models as probabilistic generative methods, and compared the results by the normal whole-image-based bag-of-features representation and the results reported in the literature.

The rest of this paper is organized as follows: In Section 2 we describe works related to Web images. In Section 3 we overview our approach and image representation, and In Section 4 we explain both generative and discriminative methods for filtering Web images. In Section 5 we presents the experimental results and evaluations, and in Section 6 we conclude this paper.

2. RELATED WORK

In the existing works on Web images, temporary training images were selected based on ranking of Google results [10], HTML analysis [30, 21], RANSAC [11] or k-Nearest-Neighbor strangeness [29]. Wang et al.[27] used images in the existing image databases on the Web such as Caltech-256 [4] and Flickr [12] as training images. Moreover, they learned text model from Wikipedia [28] and integrate visual and textual models to select relevant images from raw Web images. On the other hand, Vijayanarasimhan et al.[26] treated with learning from incomplete data as multiple instance learning. They regarded a raw image set obtained from Web image search engines for one query as a positive bag. Assuming that a returned image set includes at least one positive image, they applied a multiple instance learning method for Web image filtering.

As image representations in the recent works on Web images, the bag-of-features (BoF) representation [6] is common, since it has been proved that it has excellent ability to represent image concepts in the context of visual object categorization / recognition in spite of its simplicity. Several works which incorporate bag-of-features into Web image search/gathering task have been proposed so far [10, 15, 21, 32]. Fergus et al.[10] applied bag-of-features for filtering out irrelevant images from results of Google Image Search. They employed Probabilistic Latent Semantic Analysis (PLSA) [13] and its variants to select relevant images from Google results. Li et al.[15] proposed incremental Web image learning employing Hierarchical Dirichlet Process (HDP) [25]. Yanai [32] proposed to introduce bag-of-features and SVM into Web image gathering task where perfect training data is not available. Schroff et al.[21] also examined effectiveness of combination of bag-of-features and SVM extensively.

In all of the works mentioned above, they constructed BoF from whole images. On the other hand, in several works [33, 23, 24], region-based methods for Web images has been adopted. Yanai et al.[33] proposed the method with region segmentation and a probabilistic model based on a Gaussian mixture model (GMM), which made it possible to separate foreground regions from background regions. Regions are represented by their color, shape and texture. The proposed method is modification of the word-image translation model which annotates labels to regions [1]. Sun et al.[23, 24] detected foreground regions with the k-means clustering instead of GMM. The method proposed in this paper can be regarded as the bag-of-features version of these works.

2.1 Generative Methods

In this paper, we propose introducing latent topic models such as Probabilistic Latent Semantic Analysis (PLSA) [13] and Latent Dirichlet Allocation (LDA) [2], into Web image selection task. Recently, PLSA and LDA were applied to object recognition task as probabilistic generative models [22, 8, 18]. However, PLSA and LDA are not applied to Web images except [10]. The difference between this paper and [10] is that Fergus et al.[10] selected 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 translation model [1] and its variant for Web images [33], which employed region segmentation and a GMM-based probabilistic model. Since latent topic models such as PLSA and LDA were originally proposed for analyzing documents represented by bag-of-words, it is more appropriate for analyzing high dimensional image vectors represented by BoF than GMM. Therefore, in this paper we adopt latent topic models such as PLSA and LDA as a generative probabilistic model. A method to recognize images based on the mixture of topics obtained with PLSA has already proposed by Monay et al.[18]. Our work can be regarded as the Web image version of that work and the extension employing LDA with region-based BoF.

2.2 Discriminative Methods

As discriminative methods for Web images, support vector machine (SVM) is common, since SVM was used in the paper which proposed BoF and some open source codes of SVM such as SVM-light [14] and LIBSVM [5] are available on the Web. Schroff et al.[21] and Yanai [32] applied SVM to Web image selection tasks with BoF representation. Their works proved that SVM with soft margin worked well even under the condition where training data contains noise. However, both of them used not region segmentation but the normal bag-of-features as whole image representation.

As discriminative methods for region-based image representation, multiple instance learning (MIL) [17] is a common method. In the MIL setting, an images is represented by a set of feature vectors which is called a "bag". For MIL, positive bags and negative bags are needed to be prepared as training data. A positive bag includes more than one positive instances, while a negative bag consists of only negative instances. An image which includes objects associated with given keywords is regarded as a positive bag, since foreground regions and background regions are positive instances and negative instances, respectively. Although there are many methods to solve multiple instance problems, in this paper, we use SVM-based sparse multiple instance learning (sMIL)[3], which is modification of SVM for a multiple instance setting. In fact, [26] which is one of recent works on Web images used sMIL, although it is not region-based.

3. APPROACH

In this paper, we proposed a Web image selection with region-based bag-of-features, and examine if region-based bag-of-features is effective for automatic Web image selection where training data is noisy. In the proposed method, we perform region segmentation processing for all the images and extract bag-of-features (BoF) vectors from all the regions. To classify regions into foregrounds

and backgrounds, we introduce region-based bag-of-features. Foreground regions are expected to correspond to objects related to given keywords, and background regions do not correspond to given keywords directly. Most of the existing works adopted whole-image-based bag-of-features, and did not separate foregrounds from backgrounds.

As learning methods, we use both discriminative and generative supervised learning methods. In this paper, we use SVM-based sparse multiple instance learning (sMIL)[3] and a standard SVM as discriminative methods, and pLSA and LDA mixture models as probabilistic generative methods. Since supervised learning methods require training data, we prepare positive ones by HTML text analysis in advance and use randomly-sampled images as negative training data.

As the pre-processing before image selection performed by the proposed method, we gather Web images associated with given keywords with Web search engines and select **candidate images** which are likely to be associated with the given keywords by surrounding HTML text analysis based on simple heuristics. Particularly highly-rated images among the candidate images are selected as **initial training images** for supervised learning employed in the image selection process by visual analysis. While Schroff et al.[21] used a probabilistic HTML-analysis-based image ranker employing a naive Bayes model, we use the simple heuristics to select images by HTML analysis proposed in Yanai [30] for simplicity. The reason why we use a simple heuristics is that it does not need training data unlike the probabilistic method [21] and works well according to [30]. Since only HTML text analysis is not enough and we need visual analysis for Web image selection, we regard HTML text analysis as much less important than visual analysis which is our main objective.

3.1 Region-based Bag-of-Features

In this subsection, we describe how to generate region-based feature vectors based on the bag-of-features (BoF) model [6] briefly. The main idea of the bag-of-features model 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 [7] after adjusting the parameters so as to generate about eight regions per image on average.

The processing steps to generate region-based BoF are as follows:

1. Carry out region segmentation with the JSEG algorithm.
2. Sample 3000 patches per image randomly in the same way as [19].
3. Generate feature vectors for the sampled patches by the SIFT descriptor [16].
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.

4. CLASSIFICATION METHODS

In this section, we describe two kinds of supervised classification methods employed in the experiments: probabilistic latent topic mixture models and SVM-based sparse multiple instance learning.

4.1 Generative Models: Latent Topic Mixtures

As generative models, we use models based on probabilistic latent topic models. In this paper, we use the following two kinds of probabilistic models: Probabilistic Latent Semantic Analysis (PLSA) [13] and Latent Dirichlet Allocation (LDA) [2]. Both of them are originally unsupervised latent topic models for text analysis. The main idea is that we apply probabilistic latent models to all the regions extracted from the candidate images to get latent topics, decide "positive topics" and "negative topics" using the initial training images, and evaluate relevancy of each region extracted from images with the mixture model of positive and negative topics.

First, we apply latent topic methods to all the regions of all the candidate images and negative background images with the given number of topics, and get the probability of each topic over each region, $P(z|r)$. Next, we calculate the probability of being positive or negative regarding each topic, $P(pos|z)$ and $P(neg|z)$ using initial training images. Here, "positive topic" means that the latent topic generates foreground regions relevant to the given keywords, and "negative topic" means that the latent topic generates irrelevant regions. Finally, the probability of being positive over each candidate region, $P(pos|r)$, is calculated, which represents relevancy of each region. Based on $P(pos|I)$ which can be derived from $P(pos|r)$, the system can select images which is likely to be relevant from all the candidate images.

1. Apply the latent topic method, PLSA or LDA, with the given number of topics to the bag-of-features vectors of all the candidate regions and the negative regions, and get $P(z|r)$ where $z \in Z = (z_1, .., z_k)$ is the latent topic variable, and $r \in R = (r_1, ..., r_N)$ is a region.
2. Calculate the probability of being positive or negative over each topic, $P(pos|r)$ and $P(neg|r)$ using the initial training images which are automatically selected by the HTML analysis.
3. Calculate $P(pos|r) = \sum_{z \in Z} P(pos|z)P(z|r)$, and evaluate relevancy of each region to the given keywords.
4. Regard the maximum $P(pos|r)$ within each image as $P(pos|I)$ which represents the probability of being positive over the image.

4.1.1 PLSA

Probabilistic Latent Semantic Analysis (PLSA) [13] was originally proposed as a probabilistic model to extract latent topics from text documents. Since text documents are assumed to be represented as bag-of-words, we can apply PLSA to bag-of-features as it is. Recently several works applied PLSA to image categorization / classification task [22, 18].

The PLSA model is represented as the generative model of each word w in a document d:

$$P(w, d) = P(d) \sum_{z \in Z} P(w|z)P(z|d) \quad (1)$$

where $z \in Z = (z_1, .., z_k)$ is a latent topic variable, k is the number of topics, $d \in D = (d_1, ..., d_N)$ corresponds to a re-

gion expressed by the region-based bag-of-features, and $w \in W = (w_1, ..., w_M)$ is a visual word. The joint probability of the observed variables, w and d, is the marginalization over the k latent topics Z. The parameters are estimated by the EM algorithm. For full explanation of the PLSA model refer to [13].

4.1.2 LDA

Latent Dirichlet Allocation (LDA) by Blei et al. [2] is also a probabilistic model to detect latent topics from text documents represented by bag-of-words. It was proposed as a method to resolve a drawback of PLSA that the number of parameters in the models grows linearly with the size of the data which leads to serious overfitting. LDA models each image as a mixture over topic, where each vector of mixture proportions is assumed to have been drawn from a Dirichlet distribution.

Using the same notation in case of PLSA, the generative process is as follows:

1. Draw $\theta \sim \mathrm{Dirichlet}(\alpha_1, ..., \alpha_k)$

2. For each word $w_i \in W$

- **(a)** Draw a topic $z_i \sim \mathrm{Multinominal}(\theta)$
- **(b)** Draw a word $w_i \sim \mathrm{Multinominal}(\beta(z_i)))$ where $\beta(k)$ is a probability distribution over words corresponding to topic k

where α and β_k are model parameters for LDA and θ is a vector of topic mixture proportions. In this paper, we regard θ as $P(r|d)$.

To estimate model parameters with data, since this model involves an intractable integral, approximation computation such as variational inference and Markov Chain Monte Carlo sampling (MCMC) is needed. In our implementation, we used variational Bayes inference.

4.1.3 Compute relevancy of each image

In the next step, we calculate the probability of being positive or negative regarding each topic, $P(pos|z)$ and $P(neg|z)$ using initial training images. The probability of being positive and negative over a topic is calculated by evaluating the ratio of positive regions and negative regions. Here, "positive regions" are regions extracted from initial positive training images (PO), and "negative regions" are regions extracted from negative background training images (BG).

Calculation of $P(pos|z)$ and $P(neg|z)$ is as follows:

$$p_0 = \frac{1}{N_{PO}} \sum_{d \in PO} P(r|z) \quad (2)$$

$$p_1 = \frac{1}{N_{BG}} \sum_{d \in BG} P(r|z) \quad (3)$$

$$P(pos|z) = p_0/(p_0 + p_1) \quad (4)$$

$$P(neg|z) = p_1/(p_0 + p_1) \quad (5)$$

where,

$$P(d|z) = \frac{P(z|r)P(r)}{\sum_{r \in PO+BG} P(z|r)P(r)}, \quad (6)$$

and N_{PO} and N_{BG} represent the total number of regions in the positive images and the total number of regions in the negative images, respectively.

Finally, we can calculate the probability of being positive over each region $P(pos|r)$ by marginalization over topics:

$$P(pos|r) = \sum_{z \in Z} P(pos|z)P(z|r) \quad (7)$$

In the experiments, we set the number of topic z as 100, which was decided empirically.

We regard the maximum $P(pos|r)$ within each image as $P(pos|I)$ which represents the probability of being positive over the image.

$$P(pos|I) = \max_{d \in I} P(pos|r) \quad (8)$$

We rank all the candidate image in the descending order of $P(pos|I)$, and obtain the final result.

4.2 Discriminative Models: Multiple Instance SVM

To classify each region of candidate images as relevant and irrelevant, we also use the SVM-based sparse multiple-instance learning (sMIL) [3] and a standard SVM classifier.

Since sMIL assumes multiple instance settings, it needs positive bags, which include more than one positive instances, and negative bags, which consist of only negative instances. Then, we regard a set of the regions extracted from one of the initial positive images as a positive bag, and regard a set of the regions extracted from one of the negative images as a negative bag. For a standard SVM, we regard all the regions in the positive bags as positive regions ignoring the fact that initial positive images include noise and background regions.

The sparse MIL (sMIL)[3] is a modification of a standard SVM for the multiple instance setting. One of the constrains for optimization is adapted so that at lease one instance of positive bags should be positive. For full explanation of the sparse MIL refer to [3].

We train sMIL and standard SVM classifiers with positive and negative bags. Next, we apply the trained classifiers to all the vectors of 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.

To integrate several SVM outputs of regions in one bag into one output, we adopt simple maximum strategy which is the same as generative methods we employ in this paper. We regard the maximum SVM output values within one image as the output value of the image.

5. 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. The first ten concepts are the same as [33] which employed region-based Web image selection with GMM, while the last five concepts are commonly used in several works such as [10, 21, 29].

Before the experiments of image selection, we obtained around 5000 URLs for each concept from several Web search engines including Google Search and Yahoo Web Search. Since both search engines return 1000 URLs for one given keyword, we used the

query expansion method we proposed before [31] to obtain 5000 URLs for one concept.

After that, we selected initial training images and candidate images by using heuristic-based HTML analysis proposed in [30]. Candidate images are all the images selected by the HTML analysis, and initial training images are highly-evaluated images among the candidate images. Table 1 shows the precision of top 100 output images of Google Image Search for comparison, the number and the precision of initial training images and candidate images, and the results of image selection by the region-based probabilistic method employing a Gaussian Mixture Model (GMM) [33] for comparison. 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 [21].

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 the candidate images, 62.2%, was almost equivalent to the average precision of the top 100 results of Google images, 63.6%, while we selected about 3000 images a concept. The precision of the initial positive images was better than the precision of the Google images by about ten points because of selection of initial positive images with heuristic HTML analysis. These results show that the simple heuristic method to select initial positive images and candidate images worked well.

In the experiments, we tried seven combinations of four classifiers and two image representations: standard BoF with a pLSA mixture model (img/pLSA), standard BoF with a LDA mixture model (img/LDA), region-based BoF with a pLSA mixture model (reg/pLSA), region-based BoF with a LDA mixture model (reg/LDA). standard BoF with SVM (img/SVM), region-based BoF with SVM (reg/SVM), and region-based BoF with sMIL (reg/sMIL). Table 2 shows the results of seven combinations for the fifteen concepts.

For the first ten concepts, "img/SVM" achieved the best result, and "reg/sMIL" achieved the second best result, which outperformed the 80.0% precision by the GMM-based probabilistic method. Since "img/SVM" was very effective for the scene concepts, the result for the ten concepts including scene concepts was very good. For the six object concept in the first ten concepts, the difference between "img/SVM" and "reg/sMIL" is not so large. Although for most of the concepts the precision of "reg/sMIL" 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 "img/SVM" and "reg/sMIL", respectively.

In general, "reg/sMIL" was superior to "reg/SVM", since "sMIL" is a modification of a standard SVM to adapt for the MIL settings. The results of the ten concepts by the generative models were inferior to ones by the discriminative models in case of whole image BoF, while in case of region-based BoF "reg/pLSA" and "reg/LDA" were almost equivalent to "reg/sMIL".

For the last five concepts which is commonly used in this kind of work, "reg/sMIL" achieved the best result, 89.6%, and it was better than "img/SVM" which achieved 65.5%. Regarding generative models, region-base BoF also outperformed whole-image-based BoF. From these results, region-based BoF is effective for object concepts rather than scene concepts in general, since foreground regions corresponding to objects are important for classification.

Table 2 also shows the results of Fergus et al. [10], Schroff et al. [21] and Wnuk et al. [29]. For the five concepts available, "reg/sMIL" substantively outperformed these alternative methods. For example, the precision on "airplane" shown in Figure 4 was greatly improved, since our method tries to classify images by focusing only foreground regions, while they used the whole-image bag-of-features representation. In general, the data sources are so different that we cannot compare them strictly. However, from the fact that the method of Schroff et al.[21] which employs whole-image-based BoF and a standard SVM is essentially equivalent to "img/SVM" and the method of Fergus et al.[10] which employs BoF and pLSA is also very similar to "img/pLSA" and their results are similar to each other, respectively, we can consider that the improvements due to region-based BoF is reliable, and these results demonstrate the effectiveness of our proposed method.

For the last five concepts and "notebook PC", the results by region-based methods were improved compared to whole-image-based methods, because all of them are "object" concepts, and classification of foreground and background regions separately worked effectively. On the other hand, for "mountain" and "waterfall", the results were degraded, since the combination of foreground and backgrounds are expected to express "scene" concepts.

We show parts of the result images of "notebook PC" in Figure 1, "flower" in Figure 2, "baby" in Figure 3, "airplane" in Figure 4 "motorbike" in Figure 5, and "sunset" in Figure 6 by "reg/sMIL", respectively. As by-products of region-based methods, we can obtain estimated object regions indicated by labels in these figures.

To show more experimental results, we have prepared the Web site. The URL is as follows:
`http://mm.cs.uec.ac.jp/yanai/mir10/`

6. CONCLUSIONS

In this paper, we described a new Web image selection method employing the region-based bag-of-features representation. The combination of region-based bag-of-features and the sparse MIL classifier achieved the best result among several combinations of four classifiers and two image representations. In the experiments for the five common concept keywords, we obtained the 89.6% precision on the average, which outperformed the 65.5% precision by the whole-image-based bag-of-features and the three existing works. Especially, the results of "airplane" and "leopard" were improved greatly, since separation of foreground and background regions by the MIL setting worked effectively.

In the experiments, we found the tendency that region-based BoF was effective for "object" concepts, while standard whole-image-based BoF was effective for "scene" concepts.

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.

Table 1: The precision of top 100 output images of Google Image Search for comparison, the number and the precision (at 15% recall) of the initial positive images and the candidate images which are selected automatically by the HTML analysis, and the results of image selection by the region-based probabilistic method employing GMM [32] for comparison. The precision of the candidate images almost equals to the precision of the Google Image Search, while the precision of the initial positive images becomes better than the Google Image Search by about 10 points due to the HTML analysis. Moreover, GMM-based visual analysis improved the precision.

concepts	Google	positive	candidate	GMM [32]
sunset	85	790 (67)	1500 (55.3)	100.0
mountain	57	1950 (88)	5837 (79.2)	96.5
waterfall	78	2065 (71)	4649 (70.3)	82.0
beach	67	768 (69)	1923 (65.5)	75.0
flower	71	576 (72)	1994 (69.6)	78.5
lion	52	511 (87)	2059 (66.0)	74.6
apple	49	1141 (78)	3278 (64.3)	81.0
baby	39	1833 (56)	3571 (54.5)	70.7
notebook PC	70	781 (57)	2537 (43.6)	70.5
Chinese noodle	68	901 (78)	2596 (66.6)	70.9
TOTAL/AVG.	**63.6**	**11316 (72)**	**29944 (62.2)**	**80.0**

Table 2: The precision (at 15% recall) of the results by whole-image-based bag-of-features and pLSA/LDA(img/pLSA, img/LDA), and region-based bag-of-features and pLSA/LDA (reg/pLSA, reg/LDA), whole-image-based bag-of-features and SVM (img/SVM), region-based bag-of-features and SVM (reg/SVM), region-based bag-of-features and sMIL (reg/sMIL), and comparison with Fergus et al. [10], Schroff et al. [21], and Wnuk et al. [29] by the 15%-recall precision.

concepts	img/pLSA	img/LDA	reg/pLSA	reg/LDA	img/SVM	reg/SVM	reg/sMIL	Fergus [10]	Schroff [21]	Wnuk [29]
sunset	97.0	94.2	98.0	99.0	98.0	94.2	95.1	N/A	N/A	N/A
mountain	96.5	95.2	99.3	95.9	100.0	89.7	93.9	N/A	N/A	N/A
waterfall	70.2	54.1	84.9	86.9	96.7	86.4	90.1	N/A	N/A	N/A
beach	77.8	82.3	94.2	95.2	99.0	95.2	92.5	N/A	N/A	N/A
flower	79.9	79.0	84.9	87.7	93.8	82.3	91.8	N/A	N/A	N/A
lion	68.0	85.7	73.3	68.8	84.6	80.5	83.5	N/A	N/A	N/A
apple	82.7	67.0	94.4	94.4	93.1	79.8	88.2	N/A	N/A	N/A
baby	35.6	38.6	44.4	43.0	66.0	50.4	53.8	N/A	N/A	N/A
notebook-PC	32.8	31.3	75.0	66.0	53.7	74.2	63.5	N/A	N/A	N/A
Chinese-noodle	89.5	88.2	92.3	93.8	96.8	75.0	95.2	N/A	N/A	N/A
AVG. (4 scene)	85.4	81.4	94.1	94.2	98.4	91.3	92.9	N/A	N/A	N/A
AVG. (6 object)	64.8	65.0	77.4	75.6	81.3	73.7	79.3	N/A	N/A	N/A
AVG. (10 words)	**73.0**	**71.6**	**84.1**	**83.0**	**88.2**	**80.8**	**84.8**	N/A	N/A	N/A
airplane	46.8	44.4	68.0	67.1	58.9	83.6	92.7	57.0	45.0	76.2
guitar	63.2	64.3	75.0	79.1	63.2	88.9	77.4	50.0	72.0	80.4
leopard	76.2	87.3	90.6	85.7	46.7	100.0	98.0	59.0	72.0	58.3
motorbike	60.3	53.4	76.1	76.1	67.5	64.2	84.3	71.0	81.0	84.1
watch	89.3	93.2	95.5	94.9	91.4	90.9	95.5	88.0	97.0	88.9
AVG. (5 words)	**67.2**	**68.5**	**81.0**	**80.6**	**65.5**	**85.5**	**89.6**	**65.0**	**73.4**	**77.6**

7. REFERENCES

[1] K. Barnard, P. Duygulu, N. d. Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] R. Bunescu and R. Mooney. Multiple instance learning for sparse positive bags. In *Proc. of International Conference on Machine Learning*, 2007.

[4] Caltech 101/256 image dateset. *http://www.vision.caltech.edu/Image_Datasets/Caltech256/*.

[5] C. C. Chang and C. J. Lin. *LIBSVM: A Library for Support Vector Machines*. http://www.csie.ntu.edu.tw/c̃jlin/libsvm/.

[6] 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.

[7] Y. Deng and B. S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):800–810, 2001.

[8] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[9] H. Feng, R. Shi, and T. Chua. A bootstrapping framework for annotating and retrieving WWW images. In *Proc. of ACM International Conference Multimedia*, pages 960–967, 2004.

[10] 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.

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

[12] Flickr. *http://www.flickr.com/*.

[13] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 43:177–196, 2001.

[14] T. Joachims. *SVM* ^{light}. http://svmlight.joachims.org/.

[15] L. Li and L. Fei-Fei. OPTIMOL: automatic Online Picture collecTion via Incremental MOdel Learning. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2007.

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

[17] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *Proc. of 15th International Conference on Machine Learning*, pages 341–349, 1998.

[18] 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.

[19] 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*, pages IV:490–503, 2006.

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

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

[22] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their localization in images. In *Proc. of IEEE International Conference on Computer Vision*, pages 370–377, 2005.

[23] Y. Sun, S. Shimada, and M. Morimoto. Visual pattern discovery using Web images. In *Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 127–136, 2006.

[24] Y. Sun, S. Shimada, Y. Taniguchi, and A. Kojima. A novel region-based approach to visual concept modeling using Web images. In *Proc. of ACM International Conference Multimedia*, pages 635–638, 2008.

[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[26] S. Vijayanarasimhan and K. Grauman. Keywords to visual categories: Multiple-instance learning forweakly supervised object categorization. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2008.

[27] G. Wang and D. Forsyth. Object image retrieval by exploiting online knowledge resources. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2008.

[28] Wikipedia. *http://wikipedia.org/*.

[29] K. Wnuk and S. Soatto. Filtering internet image search results towards keyword based category recognition. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2008.

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

[31] K. Yanai. Image collector II: A system for gathering more than one thousand images from the web for one keyword. In *Proc. of IEEE International Conference on Multimedia and Expo*, volume I, pages 785–788, 2003.

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

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

Figure 1: "Notebook PC" images. The labels are added to the regions classified as positive.

Figure 2: "Flower" images. The labels are added to the regions classified as positive.

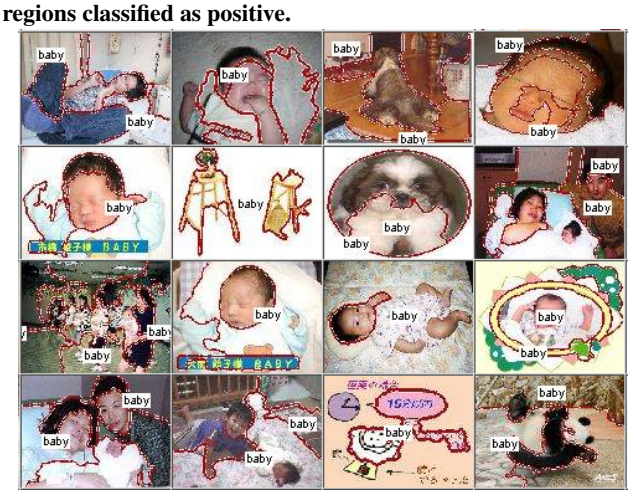

Figure 3: "Baby" images.

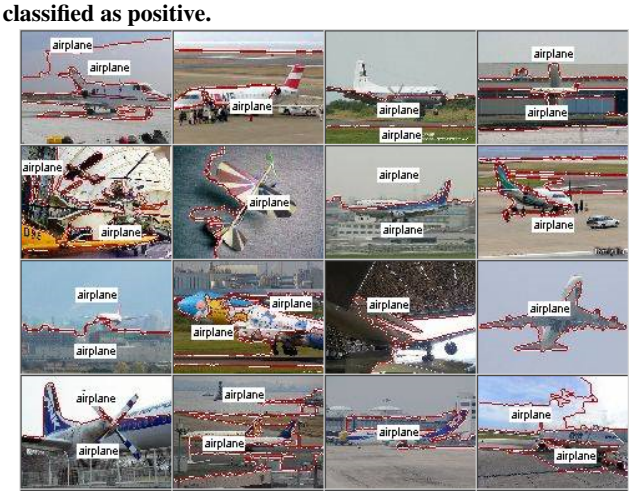

Figure 4: "Airplane" images.

Figure 5: "Motorbike" images.

Figure 6: "Sunset" images.