WEB IMAGE MINING USING CONCEPT SENSITIVE MARKOV STATIONARY FEATURES

Chunjie Zhang, Jing Liu, Hanqing Lu, Songde Ma

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, P.O. Box 2728, Beijing, China
{cjzhang, jliu, luhq}@nlpr.ia.ac.cn, mostma@gmail.com

ABSTRACT

With the explosive growth of web resources, how to mine semantically relevant images efficiently becomes a challenging and necessary task. In this paper, we propose a concept sensitive Markov stationary feature (C-MSF) to represent images and also present a classifier based scheme for web image mining. First, through analyzing the results of Google Image Searcher, we collect an image set, which are highly relevant to a concept. Then the image set is explored to learn a C-MSF about the concept by the algorithm of random walk with restart (RWR), in which the spatial co-occurrence of the bag-of-words representation and the concept information are integrated. Obtaining the concept sensitive representation, SVM is applied to mine the web images, while the highly relevant set are considered as positive examples and other random images as negative ones. Finally, experiments on a crawled web dataset demonstrate the improved performance of the proposed scheme.

Index Terms— Image mining, Image classification, feature extraction, Markov model

1. INTRODUCTION

With the development on Internet techniques and digital equipments, the number of images has exploded rapidly. To better support image indexing and retrieval at semantic level, the research on web image mining or collecting are really necessary. To enable web image mining effectively, a suitable image representation to interpret the semantics of images is one of important and challenging problems.

A lot of researchers have attempted to deal with this problem. [1] focused on the appearance and shape features by patches and curves without considering the color and texture features of images. [2] used both visual features and word vector extracted from associated HTML documents as image representation. Their co-learning algorithm required human interaction which restricted the application of their method. [3] proposed to model concept-sensitive salient regions for web image mining using Gaussian Mixture Model. However, its performance heavily depends on the usefulness of image segmentation and salient point detection. In [4], bag-of-words (BOW) based image representation was used to mine visual knowledge on the Web. The results indicate that the bag-of-words representation is much more efficient in image classification than region-based methods [3, 5]. But as a type of histogram based methods, the BOW representation has no consideration about the spatial characteristics of image features. Going one step further beyond histograms, [6] proposed the Markov Stationary Features (MSF) to involve spatial structure information of images, and employed it to obtain a unified representation for various concepts. Considering that the images with different semantics usually take on different BOW distributions, a certain concept-sensitive representation is needed to describe images effectively and further enhance the performance of web image mining.

In this paper, we propose a novel image representation called as concept sensitive Markov stationary features (C-MSF) to mine relevant images from the World Wide Web. By exploring the basic idea of the RWR algorithm, the proposed C-MSF not only extends the BOW representation with spatial structure information as in [6] but also involves the concept information of images into the final representation. With the help of Google Image Searcher, semantically relevant images are collected, and the corresponding C-MSF is extracted to represent the relevant images. A SVM classifier is then trained to mine several images on the web.

The rest of the paper is organized as follows. In Section 2, we will introduce the concept sensitive Markov stationary features (C-MSF), and in Section 3 we present classification based web image mining using C-MSF. Experimental results and conclusion are given in Section 4 and Section 5 respectively.

2. CONCEPT SENSITIVE MARKOV STATIONARY FEATURES

In this section, we will first briefly introduce the bag-of-words based image representation, and then propose our concept sensitive Markov stationary features (C-MSF).

2.1 Bag-Of-Words Based Image Representation

The bag-of-words (BOW) based image representation causes wide attention in the research community for its excellent ability and simplicity in representing image concepts. In the BOW, a visual vocabulary is generated through grouping similar keypoints into a large number of clusters and treating each cluster as a visual word. By mapping the keypoints back into the vocabulary, a histogram of visual words is constructed, which forms the feature clue to represent image content. This representation has good resistance to occlusions and within-class shape variations, and has been proven effective in many applications [4, 7]. However, it neglects the spatial structure information among the keypoints, which can provide informative knowledge to understand an image.

2.2 Concept Sensitive Markov Stationary Features

In order to utilize the spatial structure information of images, [6] proposed Markov stationary features (MSF) to extend histogram based features. For clarity, we first introduce the extraction of MSF as follows.

Let p_k be a pixel in image *I*, the spatial co-occurrence matrix is defined as $C = (c_{ij})_{K\times K}$ where

$$c_{ij} = \#(p_1 = c_i, p_2 = c_j \mid\mid p_1 - p_2 \mid = d)/2, \quad (1)$$

in which *d* (in our experiments *d=1*) indicates L_1 distance between the positions of two pixels p_1 and p_2, and c_{ij} counts the number of spatial co-occurrence for bins c_i and c_j. The co-occurrence matrix can be interpreted in a statistical view. Markov chain model is adopted to characterize the spatial relationship between histogram bins. The bins are treated as states in Markov chain models, and the co-occurrence is viewed as the transition probability between bins. In this way, the MSF extends histogram based features with spatial structure information of images. The elements of the transition matrix P are constructed from the spatial co-occurrence $C = (c_{ij})_{K\times K}$ by

$$p_{ij} = c_{ij} / \sum_{j=1}^{K} c_{ij} \quad (2)$$

If the state distribution after n steps is $u(n)$, then the Markov stationary feature (MSF) obeys

$$u(n+1) = P * u(n) \quad (3)$$

where the elements of the initial distribution $u(0)$ is defined as

$$u(0)_i = c_{ii} / \sum_{i=1}^{K} c_{ii} \quad (4)$$

According to the iterative process as Eq. 3, we can get a stable solution. Ideally, we can get a distribution of u called a stationary distribution which satisfies

$$u = P * u \quad (5)$$

The stationary distribution becomes the final representation of MSF. Obtaining the MSF of each image, the comparison of two histograms is transferred to the comparison of two corresponding Markov chains.

Although MSF has been proven effective in [6], there is still some room for improvement. Firstly, the transition matrix of MSF for one image is only affected by the spatial structure of the image. Considering the diversity of web images, this transition matrix is probably contaminated with noise. Secondly, besides the spatial structure information, images corresponding to different concepts usually take on varying feature distributions. The concept information should also be incorporated to obtain a better representation. In case of the MSF representation, the co-occurrence of visual words should also reflect the concept information. For clarity, we take an ideal example to explain it. Assuming images of 'person' include visual words of 'eye', 'hand' and 'leg', these visual words should probably appears in an image simultaneously. However, the co-occurrence for visual word 'eye' in person and 'wheel' in car is impossible to be high. By combining the different co-occurrence probability of visual words for different semantics with the spatial structure information of images, we can end up with more discriminative features for image representation than MSF.

We propose the concept sensitive Markov stationary features (C-MSF) to integrate the concept information with the spatial structure information of images. Using the same symbols as above, the C-MSF obeys

$$u(n+1) = \alpha * P * u(n) + (1-\alpha) * v \quad (6)$$

where v is the concept information vector, P is the transition matrix and α is the parameter to leverage the influence of concept information and spatial structure information.

After several iterations according to Eq. 6, the stationary distribution of u should satisfy.

$$u = \alpha * P * u + (1-\alpha) * v \quad (7)$$

which has a convergent solution as follows:

$$u = (1-\alpha)(I - \alpha * P)^{-1} v \quad (8)$$

And we define the solution in Eq. 8 as the proposed C-MSF.

In this way, the new representation of C-MSF is not only affected by the spatial structure information but also by the concept information of images. Actually, the MSF can be regarded as a special case of the proposed C-MSF, when α is set to 1. Yet, the C-MSF still keeps simplicity, compactness, and robustness.

Eq. 6 and Eq. 7 have the same formation to the random walk with restart (RWR) algorithm [8], which has been proven very effective in many applications. The addition of concept information also makes the C-MSF more resistant to noise compared with MSF. This is because the MSF is only relevant to the transition matrix calculated within one image, while the C-MSF incorporates the concept information as a prior knowledge besides the transition

matrix within one image.

The concept information is calculated using many images belonging to the same concept. We use the mean MSF of the positive training images as the concept information vector v in Eq. 6 and Eq. 7. The reason why we use the average is that it is easy to compute, resistant to noise but can represent the concept's visual information.

The novelty of the proposed C-MSF lies in two aspects. Firstly, the C-MSF extends the bag-of-words based image representation, and also owns the discriminative power as the local features which are resistant to occlusions and within-class shape variations. Secondly, it combines the concept information with the spatial structure information of images and goes one step further beyond MSF.

3. CLASSIFICATION BASED WEB IMAGE MINING USING C-MSF

The proposed system can automatically gather most of the relevant images to the keywords provided by a user. That is, the input of the system is just keywords, and the output is several hundreds or thousands of images associated with the keywords. Our proposed method consists of two stages. They are the collection stage and the selection stage sequentially. The whole process is illustrated in following Fig. 1.

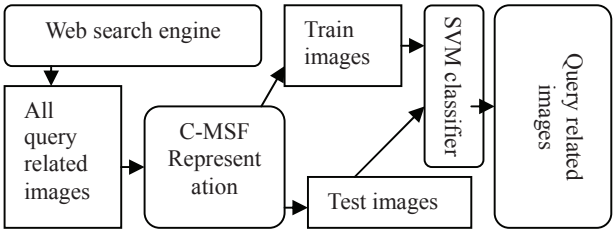

Figure 1. Flow of the proposed web images mining

3.1 Keyword Based Images Collection

Popular image search engines are becoming more and more powerful and they can provide results given a textual query. Here, we select the well-known Google Image Searcher to gather images related to a given keywords using Google Image Searcher. For the collected images, we exclude ones outside a reasonable size range (between 0.25 and 4 for the pixel ratios on both axis) and resize these images to 300×300 pixels. Then we divide these collected images into two groups according to the ranking order of Google Image Searcher. Images in group A are viewed to be highly relevant to the keywords which are used as the positive training images prepared for the following stage as mentioned in section 3.2, and the other images are classified into group B.

3.2 Classification Based Web Image Mining

In the selection stage, we train a SVM classifier to select relevant images from all the collected images. To train a classifier, we need labeled images. We could use relevance feedback to get labeled images, but this is time consuming and can also be influenced by personal preferences. In order to achieve automatic Web image mining, we regard the images in group A as positive samples, although they include a small number of irrelevant images. Images collected by other concepts are regarded as negative samples. We calculate the C-MSF of all these collected images. Then we train a SVM classifier to judge the relevance of these collected images.

4. EXPERIMENTS

We conduct experiments for the following nine concepts covering natural scenes and objects: church, dolphin, horses, hurricane, lion, mountains, sunset, tiger, and trees. For "lion" and "tiger", we add subsidiary keywords "animal" to restrict the meaning to "lion of animal" and "tiger of animal" in the collection stage respectively.

In the collection stage, we gathered around 1000 images for each concept from Google Image Searcher. This is the maximum number that Google Image Searcher provides. The number of collected images for each concept varies because of dead links. We manually identify the relevance of all the collected images as the ground truth.

In the feature extraction stage, we densely calculate the SIFT descriptors on one scale (8 orientations and 4×4 blocks of cells, with the cells being 3×3 pixels) with overlap. The overlap of patches is set to six pixels, and we ignore the 3 pixels on the edges. In this way, one image is represented by 48×48 patches of SIFT descriptors. We use k-means to cluster these SIFT descriptors into a codebook of size 600. 1-norm distance is used to compare and cluster descriptors.

We regard the first 20 images ranked by Google Image Searcher as group A. We use the images in group A as positive samples and randomly draw 2 images per concept from the other eight concepts and regard them as negative samples. We ensure the same training set is used for the three methods in our experiments. SVM classifier with RBF kernel is used in our experiment. We compare the C-MSF with MSF and BOW based image representation. To avoid randomness of choosing negative samples, we run this process for ten times and use the average of the results as the final result. We use the parameter setting so that the recall rates of C-MSF are close to the recall rates of MSF and BOW in Table 1 for easy comparison, just as [4, 5].

Table 1 shows the number, the precision and the recall of the results of BOW, MSF and C-MSF. We can not estimate the recall for the downloaded images, because the denominator to estimate it corresponds to the number of the whole Web images associated to the given concept and we can not get to know it. So we just give the precision of the raw images. However, in web image mining task, the recall rate is less important than the precision rate, since the more

Table 1. Peformance comparison among BoW, MSF and C-MSF. Numerical values out of bracket denote the number of mined images, and the ones in brackets present the precisions and the recalls respectively.

Concepts	raw images			BoW	MSF	C-MSF
	A	B	A+B	A+B	A+B	A+B
church	20 (85.0)	934 (65.1)	954 (65.5)	567 (72.6, 65.7)	593 (68.1, 65.1)	**573 (72.8, 66.6)**
dolphin	20 (95.0)	916 (59.1)	936 (59.8)	308 (62.9, 36.3)	324 (62.8, 36.5)	**276 (75.1, 37.5)**
horse	20 (95.0)	936 (76.1)	956 (76.5)	886 (77.9, 94.4)	885 (77.5, 93.8)	**866 (78.1, 92.5)**
hurricane	20 (100)	950 (32.2)	970 (33.6)	623 (42.4, 80.9)	**633 (43.1, 79.6)**	641 (42.1, 79.8)
lion	20 (100)	950 (42.2)	970 (43.4)	460 (52.9, 57.8)	462 (54.3, 57.9)	**436 (55.6, 57.7)**
mountain	20 (90.0)	893 (66.0)	913 (66.5)	291 (74.7, 35.8)	**275 (81.4, 36.9)**	295 (71.8, 35.0)
sunset	20 (95.0)	941 (65.9)	961 (66.5)	365 (84.5, 48.3)	375 (81.9, 47.5)	**338 (92.7, 48.2)**
tiger	20 (95.0)	895 (31.7)	915 (33.1)	683 (38.3, 86.3)	680 (38.0, 84.9)	**662 (40.1, 86.4)**
tree	20 (90.0)	942 (68.8)	962 (69.2)	639 (72.6, 69.6)	514 (73.8, 55.8)	**537 (85.7, 69.0)**
TOTAL/AVG.	180 (93.9)	8357 (56.3)	8537 (57.1)	4822 (64.3,63.9)	4741 (64.5, 62.0)	**4624 (68.2, 63.6)**

web sites we crawl, the more images we can get easily. So we mainly evaluate the performance of BOW, MSF and C-MSF by the precision.

For C-MSF, we obtained the 68.2% precision on the average, which outperformed the 64.3% precision by the BOW method and the 64.5% precision by the MSF method. Except "mountain" and "hurricane", the precisions of C-MSF for each concept were also improved compared with MSF, especially for "dolphin", "sunset" and "tree", which shows the effectiveness of adding concept information in representing images. The BOW, MSF and C-MSF methods all outperform the performance of Google Image Searcher in precision.

We can see from Table 1 that the C-MSF is more robust to noise than MSF, because as pointed out by [4], the MSF performs better than BOW because MSF involves spatial structure information. However, the diverse of web images make the MSF not so robust in representing images while the adding of concept information makes our C-MSF more robust to noise than MSF.

5. CONCLUSION

In this paper, we presented a classifier based scheme for web image mining and also proposed a concept sensitive Markov stationary feature (C-MSF) to represent images. By imitating the algorithm of random walk with restart, the C-MSF incorporates the concept information as a prior knowledge and the spatial co-occurrence of histogram patterns as the basic transition probability to get a new compact image representation. It achieves more robustness, simplicity, compactness than the MSF as well as the classical BOW. Experimental results demonstrated the good performance of the proposed mining scheme based on the C-MSF.

Our future work will consider the following two directions. First, we will try to prepare more and better raw group A images. Second, we plan to combine the concept information more intelligently.

6. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (Grant No.60835002, 60675003 and 60723005).

7. REFERENCES

[1] R. Fergus, P. Perona, and A. Zisserman, "A visual category filter for google images," In *Proc. Of European Conference on Computer Vision*, pages 242-256, May 2004.

[2] 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.

[3] J. Liu, Q. Liu, J. Wang, H. Lu, S. Ma, "Web image mining based on modeling concept-sensitive salient regions," *Int'l Conf. on Multimedia & Expo*, July 2006.

[4] K. Yanai, "Image collector III: A web image-gathering system with Bag-of-keypoints," *World Wide Web Conference*, pages 1295-1296, May 2007.

[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.

[6] J. Li, W. Wu, T. Wang and Y. Zhang, "One step beyond histograms: Image representation using Markov stationary features," *Computer Vision and Pattern Recognition*, June 2008.

[7] D. G. Lowe. "Distinctive image features from scale- invariant keypoints," In *Proc. Of ECCV Workshop on Statistical Learning in Computer Vision*, 60(2):91-110, 2004.

[8] L. Page et al. "The PageRank citation ranking: bring order to the web," *technical report*, Stanford Digital Library Technologies, 1999-0120, Jan. 1998.