WEB IMAGE MINING BASED ON MODELING CONCEPT-SENSITIVE SALIENT REGIONS

Jing Liu, Qingshan Liu, Jinqiao Wang, Hanqing Lu, Songde Ma

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

ABSTRACT

In this paper, we propose a probabilistic model for web image mining, which is based on concept-sensitive salient regions without human intervene. Our goal is to achieve a middle-level understanding of image semantics to bridge the semantic gap existing in the field of image mining and retrieval. With the help of a popular search engine, semantically relevant images are collected, and concept-sensitive salient regions are extracted automatically based on an attention model. Then the semantic concept model is learned from the joint distribution of all salient regions with Gaussian Mixture Model and Expectation-Maximization algorithm. In addition, by incorporating semantically irrelevant un-salient regions as negative samples, the discriminative power of the solution is further enhanced. Experiments demonstrate the encouraging performance of the proposed method.

1. INTRODUCTION

With the exponential growth of web images, it is urgent to develop image-mining technologies based on semantics in order to effectively index and retrieval images. In order to bridge the semantic gap between low-level visual features and high-level human interpretation, a middle-level semantic understanding of images becomes very important. However, how to find and represent the underlying concept model is a key issue.

Several researchers have attempted to deal with these problems. In [8], multiple instances learning algorithm was employed to build someone's face model from the result of Google Image Search. In this method, the skin detector needs to be trained in advance and the validity of the visual model deeply depends on the skin detector. Furthermore, it has a limitation that the subject must be human. [5] and [7] proposed probabilistic modeling algorithms for ranking web images relevant to specified categories. [7] focused on the appearance and shape features by patches and curves, without considering the important color and texture features for images visual appearance. [5] built the visual model depending on segmented regions, and the highly relevant and irrelevant regions were refined by an iterative selection. However, this method could not ensure the selected regions at semantic and object level. Additionally, in [5] and [7], negative samples were selected manually in advance to enhance the learned model's discriminative power, which were not at the semantic-level irrelevance to the specified concept.

In this paper, we propose a novel probabilistic model to mine relevant images from the WWW (World Wide Web) based on the concept-sensitive salient regions and un-salient regions. The whole process flow is as following Fig. 1.

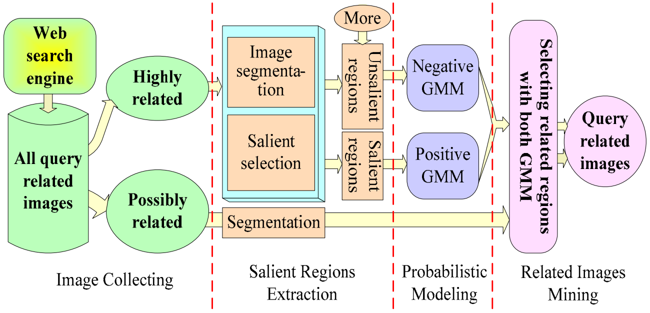

Fig. 1. Processing flow of concept relevant images collecting

The rest of this paper is organized as follows. In Section 2, we briefly introduce the method of collecting keyword-relevant images under HTML analysis. In Section 3, we present the concept-sensitive image content representation using salient regions. The probabilistic modeling method for image semantics through EM-based GMM is introduced in Section 4. Finally, experimental results and conclusion are respectively given in Section 5 and 6.

2. SEMANTIC RELEVANT IMAGES COLLECTING

Since there are abundant textual information around web images, we can easily obtain some useful semantic information for web images. Moreover, popular search engine can supply good results for a semantic-level query. Thus based on the existing search engine, we gather query relevant images by analyzing corresponding HTML documents. The details for the image gathering process are as follows.

In this paper, we use Google search engine to collect the images. Firstly, we submit a keyword that can represent the

1-4244-0367-7/06/$20.00 ©2006 IEEE

visual semantics of images. In order to restrict the keyword to have only one dominant meaning, sometimes we add some determiners, such as “tiger and animal”. Secondly, we gather all returned URL results and analyze the fetched HTML documents. Then images indicated by “IMG SRC” or “A HREF” tags in HTML documents are crawled. For these images, we exclude ones outside a reasonable size range (between 100 and 600 pixels on both axis, between 0.25 and 4 for the pixel ratios on both axis) and normalize the collected images by its height no more than 200 pixels. Finally we evaluate the relevance between images and the keyword by analyzing the HTML document. As a webpage designer, if he or she describes an image by a specific word in HTML, the word usually can express the image content from high semantic level. By this token, we collect images whose file name or ALT includes the specified keyword. These collected images make up highly relevant base (*H*) for building a probabilistic model, which is explained in section 3 and 4. All residual images as possibly relevant base (*P)*, prepare to be mined with the learned model.

3. CONCEPT-SENSITIVE IMAGE CONTENT REPRESENTATION

Most existing techniques for semantic image retrieval depend on the discriminative power of the visual features. Two approaches are widely used in the image content representation, i.e. image-based and region-based methods. The former uses global visual features, so it cannot work well for the object-based images. Although the latter tries to access the image at the object-level, its performance often does not reach human demands due to the problem of semantic object extraction and the validity of segmentation algorithm. To achieve more accurate interpretation of image semantics with visual features, we propose to extract the low-level features of salient regions in the highly relevant image collection.

3.1. Image Segmentation

Our main motivation for segmentation is to get visually homogenous regions without too much strict requirements. Here we employ a simple segmentation algorithm using the K-means clustering based on color-texture features [6]. The color-texture features are obtained by a set of Gabor filters (5 scales and 4 orientations) on items E_α, E_β and E_γ, which are obtained by producing RGB items with an optimized Gauss color model as follows:

$$\begin{bmatrix} E_\alpha \\ E_\beta \\ E_\gamma \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.31 \\ 0.19 & 0.18 & -0.37 \\ 0.22 & -0.44 & 0.06 \end{pmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}, \qquad (1)$$

Then we use the magnitude responses of 20 filters to construct a 60-dimensional feature vector. Additionally, we append the (x, y) position and the distance to the image center of every pixel into the feature vector. Then we apply PCA to reduce the features into a compact subspace representation for image segmentation.

3.2. Salient Region Detection

The concept-sensitive salient regions are defined as the dominant and representative regions of the image and are also visually attentive to users’ understanding. Since the salient regions are semantic to human beings, they can serve as a middle-level representation of image content to bridge the semantic gap. In this paper, we apply the attention model proposed by Itti L. [1-2] to help identify the most salient regions in the image. Some examples for salient regions extraction are shown in Fig. 2.

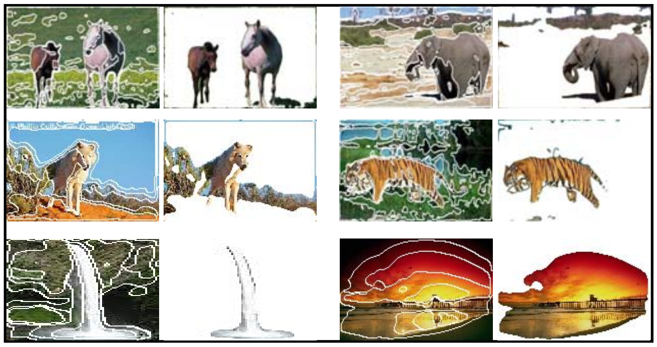

Fig. 2. The detection results of salient regions

In this model, an input image is decomposed into a set of multi-scale neural feature maps to represent local spatial discontinuities in the modalities of color, intensity and orientation. All feature maps are then combined into a unique scalar saliency map that encodes the salience of a location in the scene. Then the most salient location in the image corresponds to the locus of highest activity in the saliency map. This is achieved by using a winner-take-all neural network, which implements a neural distributed maximum detector. To allow the attention to shift to the next most salient location, each attended location is transiently inhibited in the saliency map by an inhibition-of-return mechanism, such that the winner-take-all network naturally converges towards the next most salient location. Then locations with high salient values (top n salient ones) are selected as the key salient locations. And the segmented regions including these salient locations are concept-sensitive salient regions. Other regions in an image are considered as semantically irrelevant un-salient regions, which are used as negative samples in the following model learning process.

3.3. Feature Representation

After we get the concept-sensitive salient regions, a set of visual features are calculated to characterize their principal visual properties. These features include 6-dimensional locations (i.e. 2-dimensional for region center and 4-dimensional to indicate the rectangular box for a coarse shape representation), 60-dimensional means of color-texture features for above extracted features in section 3.1.

4. PROBABILISTIC MODELING FOR MIDDLE-LEVEL PRESENTATION

To achieve the visual consistency among images relevant to a specified concept, the mixture models are used to approximate the joint distributions of all the concept-sensitive salient regions. We use the EM algorithm to estimate the parameters of the model and apply the Minimum Description Length (MDL) principle to select the number of mixture components.

4.1. EM-based Gaussian Mixture Model

Assume that we use K Gaussians in the mixture model. The probabilistic density is as follows:

$$f(x|\theta) = \sum_{i=1}^{K} \alpha_i f_i(x|\theta_i) \ , \tag{2}$$

where x is a feature vector, the α_i represents the mixing weights ($\sum_{i=1}^{K} \alpha_i = 1$), θ is the parameter collection, f_i is a multivariate Gaussian density parameterized by θ_i:

$$f_i(x|\theta_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)} \ , \tag{3}$$

where d is the dimension of the feature space.

For the value of K, we can apply the MDL principle to make decision [3-4]. Choosing K is to maximize the following expression:

$$\log L(\theta|\chi) - \frac{m_k}{2} \log N \ , \tag{4}$$

where m_k is the number of free parameters needed for a model with K mixture components. In the case of a Gaussian mixture with full covariance matrices, we have

$$m_K = (K+1) + K_d + K\frac{d(d+1)}{2} \ , \tag{5}$$

$$\log L(\theta|\chi) = \log \prod_{k=1}^{N} f(x_k|\theta) \ , \tag{6}$$

According to this principle, we can get the suitable number of components for GMM.

The EM algorithm is used for finding maximum likelihood parameter estimates when there are missing or incomplete data. In our case, the missing data is Gaussian cluster membership. We estimate values to fill in for the incomplete data (E-step), compute the maximum likelihood parameter estimates using this data (M-step), and repeat until a suitable stopping criterion is reached.

4.2. Probabilistic Model for Image Semantic Content

To get a Gaussian mixture model for concept-sensitive salient regions, we need training samples, i.e. semantically relevant regions. As section 2 mentioned, through analyzing the corresponding HTML documents, we can collect a set of highly relevant images from Internet. Then salient regions for every image in the base H can be considered as the positive training samples. In addition, we select the un-salient regions for many keyword-relevant crawled images as the negative training samples to enhance the discriminative power of the model.

In this paper, we denote the semantically relevant probabilistic model for salient regions as $P_S(R|r_j^i)$, and denote the semantically irrelevant probabilistic model for un-salient regions as $P_{\bar{S}}(\bar{R}|r_j^i)$. Then the semantically relevant probability for a region $r_j{}^i$ in image I_i and the relevant probability for image I_i are respectively as follows:

$$P(R|r_j^i) = \frac{P_S(R|r_j^i)}{P_S(R|r_j^i) + P_{\bar{S}}(\bar{R}|r_j^i)} \ , \tag{7}$$

$$P(R|I_i) = \frac{1}{T}\sum_{k=1}^{T} P(R|r_{top_k}^i) \ , \tag{8}$$

where $r_{top_k}^i$ is the i^{th} largest region within image I_i in terms of $P(R|r_j^i)$. In our experiments, we set T to 3.

Finally, we select images from the crawled collection (base H and base P), whose $P(R|I_i)$ is more than a threshold *th* as final output semantically relevant images (in our experiment *th*=0.5, i.e. the semantically relevant probability is larger than the semantically irrelevant probability).

5. EXPERIMENT

We conduct two comparison experiments for the following ten concepts covering objects and natural scenes: deer, lion, elephant, tiger, wolf, leopard, horse, sunset, mountain and waterfall.

In the image collecting stage, we gather 1000 URLs for every concept from Google Search Engine. We find three volunteers to manually identify the relevance of all the collected images, and combine their evaluations by voting. The images to be identified include highly relevant base H and possibly relevant base P, and the returned images by Google Image Search for corresponding concepts, which prepares for the following comparison experiment.

In the stage of salient region extraction, we set the dimensions of PCA subspace as 5 and the number of segmented regions as 8 for all the images. Additionally, we randomly gather around 5000 un-salient regions as negative samples from 1500 images covering these ten concepts for building the semantic unrelated model, whose component number is 48 defined by MDL principle.

To prove the effectivity of the negative samples, we compare our method with the related work by K. Yanai [5]. According to [5], we gather 1000 images by search 20 adjective keywords having no relation to above ten concepts.

After segmenting these images, we randomly select 5000 regions as the negative samples. Fig. 3 shows the precision and the number of images mined by our proposed method and Yanai's method. These values are computed by the base *H* and *P* for every concept. Regarding the results shown in Fig. 3, the prominence of our method can be proved. Maybe the randomicity of selected negative samples in [5] influences the model's discriminative power.

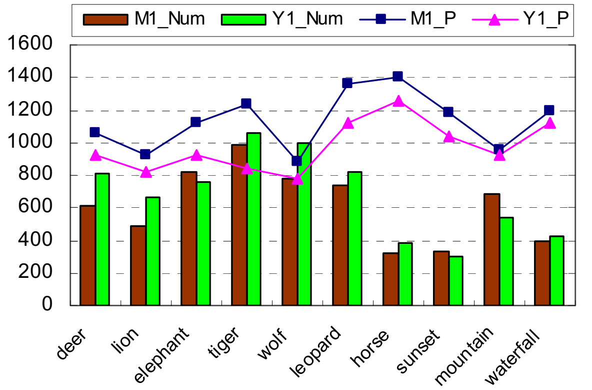

Fig. 3. The precision and the number of mined images by our method (M1) and Yanai's (Y1), wherein the lines denote the precision and bars denote the numbers according to every concept.

In Fig. 4, the comparison between our method and Google Image Search is shown. The average values of Google's GP@20, GP@100, GP@200 for ten concepts are 74%, 45.7% and 39.9% respectively, while our method denoted as MP@20, MP@100, MP@200 are 74.6%, 65.7% and 58.7% respectively. As we can see, precision of Google Image Search is more sensitive to the increasing number of returned images. However, our proposed method can achieve encouraging performance.

6. CONCLUSION

In this paper, a probabilistic modeling method is proposed for mining semantically relevant web images automatically. Through analyzing the result of Google search engine, we easily obtain the highly relevant and possibly relevant image bases. Then the concept-sensitive salient regions extracted as a middle-level understanding of image semantics are used for model learning. In addition, incorporating un-salient regions as negative samples further improves the model's discriminative power. Finally, experimental results show the improved accuracy compared with Google Image Search and K. Yanai [5]'s method.

As the surrounding text and hyperlinks are valuable high-level information for web images, our future work will focus on building more effective models for more concepts considering multi-modality combination.

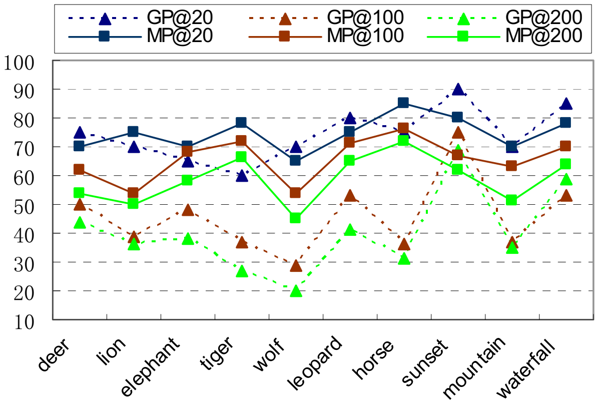

Fig. 4. The precision of top n returned images for every concept by our method (MP@20, MP@100, MP@200), compared with Google image search results (GP@20, GP@100, GP@200).

ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (Grant No. 60475050 and 60405005).

7. REFERENCES

[1] Itti L., Koch C., "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience*, 2(3), pp. 194-203, 2001.

[2] Itti L., Koch C. and Niebur E, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. PAMI*, 20(11), pp. 1254-1259, 1998.

[3] J. Rissanen, "Modeling by shortest data description," *Automatic*, 14, pp. 465-467, 1978

[4] J. Rissanen, "Stochastic complexity in statistical inquiry," *World Scientific*, 1989

[5] K. Yanai, K. Barnard, "Probabilistic Web Image Gathering," *In Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 704-707, 2005.

[6] M. Hoang, J. Geusebroek and A. Smeulders, "Color texture measurement and segmentation," *Signal Processing* 85(2), pp. 265-275, 2005.

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

[8] X. Song, C. Lin, M. Sun, "Autonomous visual model building based on image crawling through internet search engines," In Proc. of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 315-322, 2004.