Web Image Mining toward Generic Image Recognition

Keiji Yanai
Department of Computer Science, The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo, JAPAN
yanai@computer.org

ABSTRACT

We describe a generic image classification system with an automatic knowledge acquisition mechanism from the Web. The processing in the system consists of three steps. In the gathering stage, the system gathers images related to given class keywords from the Web automatically. In the learning stage, it extracts image features from gathered images and associates them with each class. In the classification stage, the system classifies an unknown image into one of the classes corresponding to the class keywords by using the association between the image features and the classes.

KEYWORDS

Web image mining, image-gathering, image classification

1. INTRODUCTION

Due to the recent wide spread of digital imaging devices, we can easily obtain digital images of various kinds of real world scenes, so that demand for image recognition of various kinds of real world images becomes greater. It is, however, hard to apply conventional image recognition methods to such generic recognition, because most of their applicable targets are restricted. Henceforth, semantic processing of images such as automatic attaching keywords to images, classification and search in terms of semantic contents of images are desired.

So far, automatic attaching keywords [1,4,6] and semantic search [2] for an image database have been proposed. In these works, since images with correct keywords were required for learning an association between images and words, commercial image collections were used for learning, for example, Corel Image Library. However, most of images in commercial image collections are well-arranged images taken by professional photographers, and many similar images are included in them. They are different from images of real world scenes taken by the people with digital cameras.

In this paper, we propose utilizing images gathered from the Web for learning of a generic image classification system instead of commercial image collections. In other words, this research is Web image mining for generic image classification. We can easily extract keywords related to an image on the Web (Web image) from the HTML file linking to it, so that we can regard a Web image as an image with related keywords. Web images are as diverse as real world scene, since Web images are taken by a large number of people for various kinds of purpose.

The processing in our system consists of three steps. In the gathering stage, the system automatically gathers images related to given class keywords from the Web. In the learning
stage, it extracts image features from gathered images and associates them with each class. In
the classification stage, the system classifies an unknown image into one of the classes
corresponding to the class keywords by using the association between image features and
classes. The system is constructed as an assembly of three modules, which are an image-
gathering module, an image-learning module, and an image classification module (Figure 1).

![Proposed System](image)

**Figure 1: Proposed System.**

2. IMAGE GATHERING

At first, we need to decide some class keywords, which represent classes into which unknown
images are classified. For example, "cow", "dog" and "cat". For each class keyword, we
gather related images from the Web. To gather images from the Web, we use the Image
Collector [10] we proposed previously as an image-gathering module.

In the gathering stage, an image-gathering module gathers images from the Web related to the
class keywords. Note that our image-gathering module is not called image "search" but image
"gathering", since it has the following properties: (1) it does not search for images over the
whole Web directly, (2) it does not make an index of the Web images in advance, and (3) it
makes use of search results of commercial keyword-based search engines for the class
keywords. These properties are different from conventional Web image search systems such
as WebSeer [5], WebSEEk [9] and Image Rover [8]. These systems search for images based
on the query keywords first, and then a user selects query images from their search results.
These three systems carry out their search in such an interactive manner. Our module is
different from those in that our system only needs one-time input of query keywords due to
automatic image selection mechanism. The details are described in [10].

3. LEARNING AND CLASSIFICATION

In the system, image classification is performed by image-feature-based search. First, in the
learning stage, an image-learning module extracts image features from gathered images and
associates image features with the classes represented by the class keywords. Next, in the
classification stage, we classify an unknown image into one of the classes by comparing
image features.

In our method of image classification, image features of not only a target object but also non-
target objects such as the background are used as a clue of classification, since non-target
objects usually have strong relation to a target object. For example, a cow usually exists with a grass field and/or a fence in a farm, and a lion usually exists in a savanna or a zoo. Although the number of combinations of a target object and non-target objects is large, we think that we can deal with this largeness by gathering a large amount of image from the Web and by using them for learning. Here, we do not set up "reject", and then all test images are classified into any class.

In the experiments, we used two kinds of image features for learning and classification: color signatures and region signatures. A signature describes multi-dimensional discrete distribution, which is represented by a set of vectors and weights. In case of color signatures, a vector and a weight correspond to a mean color vector of each cluster and its ratio of pixels belonging to that cluster, respectively, where some color clusters are made in advance by clustering color distribution of a whole image. In case of region signatures, a set of feature vectors of regions and their ratio of pixels represents a region signature, where regions are made by an image segmentation method in advance. To compute dissimilarity between two signatures, Earth Mover's Distance (EMD) has been proposed [7].

In the classification stage, we employ the k-nearest neighbor (k-NN) method to classify an unknown input image into a certain class. The value of k is decided as 5 by preliminary experiments.

4. EXPERIMENTAL RESULTS

We made six kinds of classification experiments from no.1 to no.6 for 10 kinds of gathered images, 10 kinds of gathered images with only correct ones (selected by hand), 10 kinds of images selected from the commercial image database (Corel Image Gallery), 20 kinds of gathered images, 20 kinds of images with only correct ones, and 50 kinds of images, respectively. In the experiments, we exploited three kinds of image features, which are color signatures, region signatures using the k-means clustering method, and region signatures using the JSEG segmentation algorithm [3].

In the experiment no.1, we gathered images from the Web for 10 kinds of class keywords related to animals shown in Table 1. The total number of gathered image was 4582, and the precision(pri.) by subjective evaluation was 68.2%, which is defined to be \( \frac{N_{ok}}{N_{ok}+N_{ng}} \), where \( N_{ok}, N_{ng} \) are the number of relevant images and the number of irrelevant images to their keywords.
Table 1: Results of image-gathering (left) and classification (right) in experiment no.1, no.2, and no.3.

<table>
<thead>
<tr>
<th>Class</th>
<th>num.</th>
<th>prec.</th>
<th>F</th>
<th>exp. no. 1</th>
<th>num.</th>
<th>prec.</th>
<th>F</th>
<th>exp. no. 2</th>
<th>num.</th>
<th>prec.</th>
<th>F</th>
<th>exp. no. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>bear</td>
<td>419</td>
<td>56.4</td>
<td>21.0</td>
<td>31.1</td>
<td>25.1</td>
<td>17.1</td>
<td>46.2</td>
<td>25.0</td>
<td>36.0</td>
<td>62.1</td>
<td>45.6</td>
<td></td>
</tr>
<tr>
<td>cat</td>
<td>354</td>
<td>62.0</td>
<td>28.0</td>
<td>60.9</td>
<td>38.4</td>
<td>34.3</td>
<td>78.7</td>
<td>47.8</td>
<td>61.2</td>
<td>85.7</td>
<td>71.4</td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>570</td>
<td>75.7</td>
<td>40.0</td>
<td>23.3</td>
<td>29.4</td>
<td>58.6</td>
<td>21.5</td>
<td>31.4</td>
<td>24.0</td>
<td>75.0</td>
<td>36.4</td>
<td></td>
</tr>
<tr>
<td>elephant</td>
<td>506</td>
<td>65.5</td>
<td>25.0</td>
<td>23.1</td>
<td>24.0</td>
<td>25.0</td>
<td>32.1</td>
<td>28.1</td>
<td>68.0</td>
<td>69.4</td>
<td>68.7</td>
<td></td>
</tr>
<tr>
<td>fish</td>
<td>275</td>
<td>89.9</td>
<td>22.0</td>
<td>74.6</td>
<td>31.0</td>
<td>35.7</td>
<td>62.5</td>
<td>45.5</td>
<td>58.0</td>
<td>93.5</td>
<td>71.6</td>
<td></td>
</tr>
<tr>
<td>lion</td>
<td>504</td>
<td>77.0</td>
<td>45.0</td>
<td>25.2</td>
<td>32.3</td>
<td>47.9</td>
<td>35.1</td>
<td>40.5</td>
<td>82.0</td>
<td>77.4</td>
<td>79.6</td>
<td></td>
</tr>
<tr>
<td>penguin</td>
<td>576</td>
<td>57.0</td>
<td>33.5</td>
<td>29.0</td>
<td>31.1</td>
<td>47.9</td>
<td>27.3</td>
<td>34.8</td>
<td>50.0</td>
<td>42.4</td>
<td>45.9</td>
<td></td>
</tr>
<tr>
<td>sheep</td>
<td>347</td>
<td>64.0</td>
<td>13.0</td>
<td>34.2</td>
<td>18.8</td>
<td>17.1</td>
<td>36.4</td>
<td>23.3</td>
<td>80.0</td>
<td>46.0</td>
<td>58.4</td>
<td></td>
</tr>
<tr>
<td>tiger</td>
<td>405</td>
<td>68.7</td>
<td>24.0</td>
<td>32.2</td>
<td>27.5</td>
<td>10.7</td>
<td>60.0</td>
<td>18.2</td>
<td>72.0</td>
<td>69.2</td>
<td>70.6</td>
<td></td>
</tr>
<tr>
<td>whale</td>
<td>592</td>
<td>72.4</td>
<td>66.5</td>
<td>39.0</td>
<td>49.2</td>
<td>75.0</td>
<td>55.6</td>
<td>63.8</td>
<td>94.0</td>
<td>53.4</td>
<td>68.1</td>
<td></td>
</tr>
<tr>
<td>total/avg</td>
<td>4582</td>
<td>68.2</td>
<td>31.8</td>
<td>37.3</td>
<td>34.3</td>
<td>36.9</td>
<td>45.5</td>
<td>40.8</td>
<td>62.5</td>
<td>67.4</td>
<td>64.9</td>
<td></td>
</tr>
<tr>
<td>avg.(r1)</td>
<td></td>
<td></td>
<td>29.4</td>
<td>30.3</td>
<td>29.8</td>
<td>35.4</td>
<td>37.2</td>
<td>36.2</td>
<td>67.1</td>
<td>69.2</td>
<td>68.1</td>
<td></td>
</tr>
<tr>
<td>avg.(r2)</td>
<td></td>
<td></td>
<td>24.3</td>
<td>25.4</td>
<td>24.8</td>
<td>28.4</td>
<td>28.5</td>
<td>28.4</td>
<td>47.1</td>
<td>47.7</td>
<td>47.4</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 shows the classification results evaluated by 5-fold cross-validation in the experiment no.1, no.2, and no.3. It describes only the results by color signatures for the individual classes, since most of the results by color signatures are superior to the results by region signatures using k-means and JSEG. In the tables, "(r1)" and "(r2)" mean region signatures using the k-means clustering and region signatures using the JSEG region segmentation method, respectively. The recall(rec.) is defined to be \(\frac{M_{OK}}{M_{test}}\), the precision(pri.) is defined to be \(\frac{M_{OK}}{M_{OK} + M_{NG}}\) and F-measure(F) is the harmonic mean of the recall and the precision, where \(M_{OK}\), \(M_{NG}\), and \(M_{test}\) are the number of correctly classified images, the number of incorrectly classified images, and the number of test images for each class, respectively. All values are represented in percentage terms. In the experiment no.1, we obtained 34.3 as the F-measure value by color signatures.

In the experiment no.2, we select only correct images for each class from gathered images by hand, and the classification experiment was carried out using them by 5-fold cross-validation. Compared to no.1, the F-measure increased. Especially, the result of "whale" was good, since most of "whale" images on the Web were ones of "whale watching" scene.

In the experiment no.3, we made a classification experiment not for Web images but for the 500 images of 10 classes in Corel Image Gallery as a control experiment. Since Corel Image Gallery includes many similar images to each other, a high F-measure value, 68.1, was obtained by region signatures with the k-means method.

In the experiment no.4 and no.5, we made experiments for 20 class keywords (Table 2) which include many different kinds of words. Figure 2 shows part of the images gathered from the Web. We obtained 42.3 and 46.7 as the F-measures shown in Table 3. These results are superior to the result of the experiment no.1 and no.2 for only 10 classes, because all classes used in no.1 and no.2 are related to animals and their images include many similar images even between different classes.

Table 2: 20 class keywords.

| apple, bear, bike, lake, car, cat, entrance ceremony, house, Ichiro, Ferris wheel, lion, Moai, |
Table 3: Results of experiment no.4, no.5 and no.6.

<table>
<thead>
<tr>
<th></th>
<th>exp. no.4</th>
<th>exp. no.5</th>
<th>exp. no.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg. by color</td>
<td>39.8 50.7</td>
<td>38.5 60.4</td>
<td>34.2 49.0</td>
</tr>
<tr>
<td>avg. by reg. (r1)</td>
<td>40.1 43.1</td>
<td>42.1 47.9</td>
<td>27.6 28.6</td>
</tr>
<tr>
<td>avg. by reg. (r2)</td>
<td>35.9 37.9</td>
<td>39.3 44.1</td>
<td>22.2 22.7</td>
</tr>
</tbody>
</table>

In the experiment no.6, we made a classification experiment for 50 class keywords. We obtained 40.3 as the F-measure by color signatures. This result is comparable to the results of the experiment of 20 classes. This indicates that the difficulty of classification depends on the dispersion of image features of each class in the image feature space, not simply on the number of classes.

5. CONCLUSIONS

The results in all the experiments are comparable to conventional works of generic image recognition. However, unlike them, we provide images for learning not by hand, but by gathering images related to the class keywords from the Web automatically. Although it was a troublesome task to collect such various kinds of images as ones used in the experiment no.6 so far, Web image gathering changes it into an easy task. By only providing class keywords at first without human's hands during the processing, we obtained about 40 F-measure. This shows the effectiveness of our novel method.

6. REFERENCES


