An Experiment on Generic Image Classification Using Web Images

Keiji Yanai

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

Abstract. In this paper, we describe an experiment on generic image classification using a large number of images gathered from the Web as learning images. The processing consists of three steps. In the gathering stage, a 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 a test image into one of classes corresponding to the class keywords by using the association between image features and classes. In the experiments, we achieved a classification rate 44.6% for generic images by using images gathered from the World-Wide Web automatically as learning images.

1 Introduction

Due to the recent spread of digital cameras, 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. Therefore, at present, it is impossible to deal with semantics of images of real world scene automatically. Henceforth, it is desired that automatic attaching keywords to images, classification and search in terms of semantic contents of images.

So far, automatic attaching keywords[1] and semantic search[2] for an image database have been proposed. In these works, since learning images with correct keywords were required, commercial image collections were used as learning images, 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 WWW (World-Wide Web) as learning images for generic image classification 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.

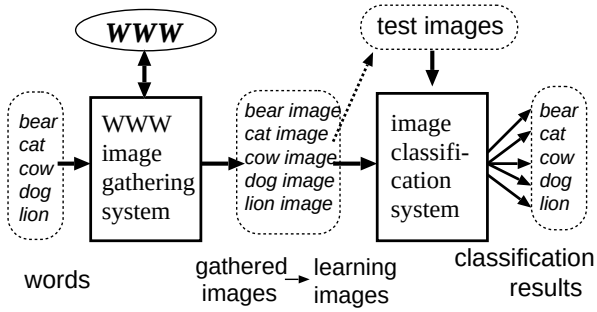

Fig. 1. Image classification by Web images.

Image classification by Web images is performed by combination of an image gathering system and an image classification system (Fig.1). First, we gather images related to some kinds of words from the Web by utilizing the Image Collector, which we have proposed in [3]. Next, we extract image features from gathered images and associate image features with words for image classification. Finally, we classify an image into one of classes corresponding to class keywords by comparing its image features with ones of images gathered from the Web in advance.

In this paper, we describe image gathering from the Web, learning and classification. Next, we describe experimental results and conclusions.

2 A Method of Image-Gathering

An image-gathering system gathers images from WWW related to the keywords given by a user. Note that our system is not called an image "search" system but an image "gathering" system, since our system has the following properties: (1) it does not search for images over the whole WWW directly, (2) it does not make a index of the Web images in advance, and (3) it makes use of search results of commercial keyword-based search engines for query keywords. These properties are different from conventional Web image search systems such as WebSeer[4], WebSEEk[5] and Image Rover[6]. These systems search for images based on the query keywords, and then a user selects query images from their search results. These three systems carry out their search in such an interactive manner. Our system is different from those in that our system only needs one-time input of query keywords due to automatic image selection mechanism described later.

Since an image on WWW is usually embedded in an HTML document that explains it, the system exploits some existing commercial keyword-based WWW search engines, and it gathers URLs (Universal Resource Locator) of HTML documents related to query keywords. In the next step, using those gathered URLs, the system fetches HTML documents from WWW, analyzes them, and evaluates the intensity of relation between the keywords and images embedded in HTML documents. If it is judged that images are related to keywords, the

image files are fetched from WWW. According to the intensity of relation to the keywords, we divide fetched images into two groups: images in group A having stronger relation to the keywords, and others in group B. For all gathered images, image features are computed. We use a color histogram in the Lu^*v^* color space as image features.

In content-based image retrieval (CBIR), a user provides query images or sketches to the system, because it searches for images based on the similarity of image features between query images and images in an image database. In our image-gathering system, instead of providing query images or sketches, a user only needs to provide query keywords to the system. Then, we select images strongly related to the keywords as group A images, remove noise images from them, and regard them as query images only by examining keywords. Removing noise images is carried out by eliminating images which belong to relatively small clusters in the result of image-feature-based clustering for group A images. Images which are not eliminated are regarded as appropriate images to the query keywords, and we store them as output images. Next, we select images that are similar to the query images from group B in the same way as CBIR, and add them to output images. The detail is described in [3].

3 A Method of Learning and Classification

We make experiments on image classification for images gathered from the Web by image-feature-based search. First, we extract image features from gathered images and associate image features with classes represented by keywords in the learning stage. Next, we classify an image into one of classes corresponding to class keywords by comparing image features in the classification stage.

3.1 Signatures and Earth Mover's Distance

We exploit two kinds of image features for learning and classification: *color signature for block segments*, and *region signature for region segments*. 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 an image. Since the number of elements of a signature is variable, it is superior to conventional fixed-size color histograms in terms of expressiveness and efficiency. In case of *region signatures*, a set of feature vectors of regions and their ratio of pixels represents a region signature.

To compute dissimilarity between two signatures, Earth Mover's Distance(EMD) has been proposed[7]. Intuitively, given two signatures, one can be seen as a mass of earth properly spread in the feature space, the other as a collection of holes in the same space. Then, the EMD measures the least amount of work needed to fill the holes with earth. Here, a unit of work corresponds to transporting a unit of earth by a unit of ground distance which is a distance in the feature space.

The EMD is based on the transportation problem and can be solved efficiently by linear optimization algorithms.

Formally, let $P = \{(\mathbf{p}_1, w_{p_1}), ..., (\mathbf{p}_m, w_{p_m})\}$ be the first set with m elements, where $\mathbf{p}_i$ is the feature vector and w_{p_i} is its weight; $Q = \{(\mathbf{q}_1, w_{q_1}), ..., (\mathbf{q}_n, w_{q_n})\}$ the second set with n elements; and $d_{ij} = d(\mathbf{p}_i, \mathbf{q}_j)$ the ground distance matrix where d_{ij} is the distance between $\mathbf{p}_i$ and $\mathbf{q}_j$. The EMD between sets P and Q is then

$$\mathrm{EMD}(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \tag{1}$$

where $\mathbf{F} = [f_{ij}]$, with $f_{ij} \geq 0$ the flow between $\mathbf{p}_i$ and $\mathbf{q}_j$, is the optimal admissible flow from P to Q. In addition, an easy-to-compute lower bound for the EMD between signatures with equal total weights is the distance between their centers of mass.

3.2 Color Signatures

To obtain *color signatures*, first, we normalize the size of learning images into 240×180, and divide them into 16 and 9 block regions as shown in Fig. 2. We make a color signature for each of these 25 block regions. Next, we select some dominant colors by clustering color vectors of each pixel into color clusters by the k-means method. In the experiments, the number of color clusters is 15 or less, and it is decided in order not to make a cluster whose weight is less than 0.005. We make a color signature for each block with elements consisting of a mean color vector of each cluster and its ratio of pixels belonging to that cluster. A mean color vector is represented by the Lu^*v^* color space which is designed in order that Euclid distance between two points in this space matches the human color sense, so that we use Euclid distance as ground distance.

In the classification stage, first, we extract color signatures from each block in an image to be classified (a test image) in the same way as the learning stage after normalizing its size. We obtain 25 sets of signatures for one test image. Next, we search all blocks of learning images of each class for the block with the minimum distance (dissimilarity) to each block of the test image. Here, the distance is computed by the EMD. In the next step, we sum up the minimum distances between the test image and learning images of each class for 25 all blocks. This search and computation is carried out for all the classes. We compare the total distances among all the classes, and we classify the test image into the class whose total distance is the smallest. In the actual implementation, we used lower bound of the EMD to reduce a frequency of computation of the EMD.

3.3 Region Signatures

To obtain *region signatures*, we carry out region segmentation for images instead of dividing images into block segments after normalizing their size (Fig.3). Many methods of region segmentation have been proposed so far. Here, we employ a simple segmentation method based on k-means clustering used in [8]. First, we

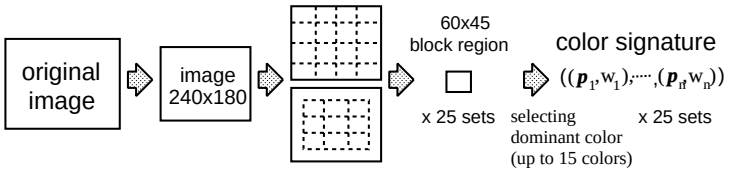

Fig. 2. Color signatures for color segments.

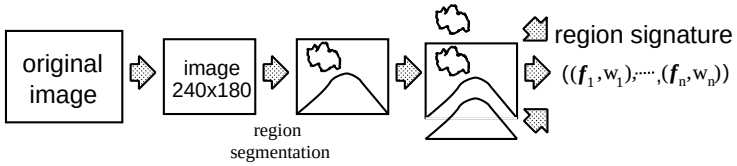

Fig. 3. Region signatures for region segments.

divide a learning image into 4×4 small blocks, and for each block we compute a mean color vector in the Lu^*v^* color space and a texture feature vector, which consists of square means of HL elements, LH elements and HH elements obtained by Daubechies-4 wavelet transform to each 4×4 block. Both vectors are three-dimension, so that a six-dimension feature vector is obtained for each block. Next, we cluster all blocks in a learning image into some regions by the k-means method in the similar way as computing color signatures. In the experiments, the number of color clusters is 15 or less, and it is decided in order not to make a cluster whose weight is less than 0.005. Then, we compute a mean 6-dimension feature vector for each region. In addition, for making a region signatures we extract three more features about shape of a region. We use normalized inertia of order 1 to 3 as three features to describe shape of a region. Finally, we make a region signature with elements consisting of a nine-dimensional feature vector for each region and its ratio of pixels belonging to that region.

In the classification stage, we employ the k-nearest neighbor (k-NN) method to classify a test image into a class. The value of k is decided as 5 by preliminary experiments. We used Euclid distance as ground distance to compute the EMD.

In our method of image classification, image features of not only a target object but also non-target objects such as background included in the image 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 grass field and/or fence in farm, and a lion usually exists in Savannah or zoo. Although the number of combination 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 using them as learning images. Here, we do not set up "reject", and then all test images are classified into any class.

Table 1. Nine experiments.

no.	num.of classes	num.of images	precision (%)	test images num.	src.
1	10	4548	68.2	20	WWW
2	10	3102	100†	20	WWW
3	10	500	100‡	10	Corel
4	10	4548	68.2	50	Corel
5	10	3102	100	50	Corel
6	20	5694	61.2	20	WWW
7	20	3485	100†	20	WWW
8	20	5694	61.2	20	special
9	20	3485	100†	20	special

†selection of correct images by hand.
‡Corel Image as a learning set.

Table 2. Results of image-gathering (left) and classification (right) in experiment no.1

class	num.	pre.	rec.	pre.	F
bear	419	56.4	21.0	31.1	25.1
cat	354	62.0	28.0	60.9	38.4
dog	570	75.7	40.0	23.3	29.4
elephant	506	65.5	25.0	23.1	24.0
tropical fish	275	89.9	22.0	74.6	34.0
lion	504	77.0	45.0	25.2	32.3
penguin	576	57.0	33.5	29.0	31.1
sheep	347	64.0	13.0	34.2	18.8
tiger	405	68.7	24.0	32.2	27.5
whale	592	72.4	66.5	39.0	49.2
total/avg.	**4582**	**68.2**	31.8	37.3	**34.3**
by region sig.			29.4	30.3	**29.8**

4 Experimental Results

We made nine experiments from no.1 to no.9 shown in Table 1.

In the experiment no.1, we gathered images from the Web for 10 kinds of words related to animals shown in Table 2. In the image-gathering processing, about ten thousands URLs were fetched from six major text search engines, Google, InfoSeek, Excite, Lycos, InfoNavi and Goo Japan. The total number of gathered image was 4582, and the precision by subjective evaluation was 68.2%, which is defined to be $N_{OK}/(N_{OK} + N_{NG})$, where N_{OK}, N_{NG} are the number of relevant images and the number of irrelevant images to their keywords. In the left side of Table 2, we show the number and the precision of gathered images.

In the image classification experiment, we regard each of the 10 words as one class. In the right side of Table 2, we show the classification result evaluated by 10-fold cross-validation. In this section, tables describe results by color signatures mainly, and results by region signatures are shown only in the bottom line of each table. In the table, the recall is defined to be M_{OK}/M_{test}, the precision is defined to be $M_{OK}/(M_{OK} + M_{NG})$ and F-measure is the harmonic mean of the recall and the precision, where M_{OK}, M_{NG}, 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. In the experiment no.1, we obtained 31.0 as the F-measure value.

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. The result is shown in Table 3. Compared to no.1, the F-measure increased. Especially, the result of "whale" was good, since most of "whale" images on the Web were images 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 picked up from Corel Image Gallery. The classification result evaluated by 10-fold cross-validation is shown in Table

Table 3. Results of image-gathering and classification in experiment no. 2, 3, 4, 5

	exp. no.2			exp. no.3			exp. no.4			exp. no.5		
class	rec.	pre.	F	rec.	pre.	F	rec.	pre.	F	rec.	pre.	F
bear	17.1	46.2	25.0	36.0	62.1	45.6	8.0	15.4	10.5	4.0	40.0	7.3
cat	34.3	78.7	47.8	61.2	85.7	71.4	4.1	33.3	7.3	6.1	42.9	10.7
dog	58.6	21.5	31.4	24.0	75.0	36.4	24.0	14.8	18.3	58.0	21.3	31.2
elephant	25.0	32.1	28.1	68.0	69.4	68.7	34.0	34.7	34.3	16.0	25.8	19.8
tropical fish	35.7	62.5	45.5	58.0	93.5	71.6	22.0	61.1	32.4	30.0	46.9	36.6
lion	47.9	35.1	40.5	82.0	77.4	79.6	30.0	19.5	23.6	36.0	27.3	31.0
penguin	47.9	27.3	34.8	50.0	42.4	45.9	26.0	19.7	22.4	48.0	25.5	33.3
sheep	17.1	36.4	23.3	80.0	46.0	58.4	8.0	23.5	11.9	4.0	18.2	6.6
tiger	10.7	60.0	18.2	72.0	69.2	70.6	4.0	7.4	5.2	10.0	45.5	16.4
whale	75.0	55.6	63.8	94.0	53.4	68.1	86.0	32.6	47.3	86.0	40.6	55.1
avg. by color	36.9	45.5	**40.8**	62.5	67.4	**64.9**	24.6	26.2	**25.4**	29.8	33.4	**31.5**
avg. by region	35.4	37.2	**36.2**	67.1	69.2	**68.1**	23.2	20.7	**21.9**	26.0	22.8	**24.3**

3. Since Corel Image Gallery includes many similar images to each other, a high F-measure value, 68.1, was obtained by region signatures.

In the experiment no.4 and no.5, we used the gathered images in the experiment no.1 and no.2 as learning images and the Corel images as test images. The results are shown in Table 3. In no.4 and no.5, we obtained 25.4 and 31.5 as F-measure, respectively. Since "dog", "tropical fish", "lion", "penguin" and "whale" have some typical patterns and both of the gathered images and the Corel images include the images with the typical patterns, their F-measure achieved high values. On the other hand, since "bear", "cat", "elephant", "sheep" and "tiger" had no typical patterns, their F-measures were relatively low.

In the experiment no.6 and no.7, we made an experiment for 20 words (Table 4) which includes many different kinds of words in the same way as the experiment no.1 and no.2. Compared to the expected F-measure, 5.0, in case of the random classification, we obtained much better F-measure, 42.3 and 46.7 shown in Table 5. 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 learning images include many similar images even between different classes. In case of "apple", "Kinkaku Temple" and "noodle", their result were about 60.0, since their scene have some typical patterns and many of their images were applicable to them. On the other hand, for "house" we obtained only a very low F-measure value, since "house" images had much variation. From these results, difficulty of classification depends on the properties of the class.

It is hard to collect such various kinds of images as images used in the experiment no.6 and no.7 by means of commercial image databases, and it has come to be possible only by image-gathering from the World-Wide Web.

In the experiment no.8 and no.9, we used the gathered images in the experiment no.6 and no.7 as learning images and a special test image set as test images. We make a special test image set by selecting various kinds of 50 typical images

Table 4. 20 class keywords

20: apple, bear, bike, lake, car, cat, entrance ceremony, house, Ichiro, Ferris wheel, lion, Moai, Kinkaku Temple, note PC, bullet train, park, penguin, noodle, wedding, Mt.Yari

Table 5. Results of experiment no.6, 7, 8, 9

	exp. no.6			exp. no.7			exp. no.8			exp. no.9		
class	rec.	pre.	F	rec.	pre.	F	rec.	pre.	F	rec.	pre.	F
avg. by color	34.9	53.6	**42.3**	35.7	67.8	**46.7**	39.8	50.7	**44.6**	38.5	60.4	**47.0**
avg. by region	34.3	37.7	**35.9**	37.0	45.5	**40.8**	40.1	43.1	**41.5**	42.1	47.9	**44.8**

for each class from Corel Image Gallery and Web images by hand. The classification results are shown in Table 5. In no.8 and no.9, we obtained 44.6 and 47.0 as F-measure, respectively. These results are comparable to conventional works of generic image recognition. However, unlike them, we provide learning images not by hand, but by gathering images from the World-Wide Web automatically.

5 Conclusions

In this paper, we described experiments on generic image classification using images gathered from the World-Wide Web. While the main targets of the conventional works on knowledge retrieval from the Web are numeric data and text data, we have proposed knowledge retrieval of image data from the Web in this paper. For future works, we plan to make much improvement in classification methods and extraction of image features to obtain more improved classification rate.

References

1. Barnard, K., Forsyth, D.A.: Learning the semantics of words and pictures. In: Proc. of IEEE International Conference on Computer Vision. Volume II. (2001) 408–415
2. Belongie, S., Carson, C., Greenspan, H., Malik, J.: Recognition of images in large databases using a learning framework. Technical Report 07-939, UC Berkeley CS Tech Report (1997)
3. Yanai, K.: Image collector: An image-gathering system from the World-Wide Web employing keyword-based search engines. In: Proc. of IEEE International Conference of Multimedia and Expo. (2001) 704–707
4. Framkel, C., Swain, M.J., Athitsos, V.: WebSeer: An image search engine for the World Wide Web. Technical Report TR-96-14, University of Chicago (1996)
5. Smith, J.R., Chang, S.F.: Visually searching the Web for content. IEEE Multimedia **4** (1997) 12–20
6. Sclaroff, S., LaCascia, M., Sethi, S., Taycher, L.: Unifying textual and visual cues for content-based image retrieval on the World Wide Web. Computer Vision and Image Understanding **75** (1999) 86–98
7. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision **40** (2000) 99–121
8. Wang, J.Z., Li, J., Wiederhold, G.: SIMPLIcity: semantics-sensitive integrated matching for picture libraries. IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001) 947–963