Web Image Mining: Can We Gather Visual Knowledge for Image Recognition from the Web ?

Keiji Yanai
yanai@cs.uec.ac.jp
Department of Computer Science, The University of Electro-Communications

Abstract

Because of the wide spread of digital imaging devices and the World Wide Web, we can easily obtain digital images of various kinds of real world scenes. Currently, however, classification/recognition of generic real world images is far from practical due to a diversity of real world scenes.

To deal with such diversity, we have proposed gathering real world images from the World-Wide Web and using them as training images for image classification. We call this research project "Web Image Mining". 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. It is expected that diverse training images enable us to classify/recognition diverse real world images. In this paper, we describe our ongoing project, "Web Image Mining for Generic Image Recognition".

1 Introduction

Due to the recent spread of digital imaging devices such as digital cameras, we can easily obtain digital images of various kinds of real world scenes. Therefore, the demand for generic image recognition/classification of various kinds of real world images becomes greater.

So far, automatic attaching keywords [1, 4, 6, 12] and semantic search [2] for an image database have been proposed. In these works, since training images with correct keywords were required, commercial image collections were used as training 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 commodity digital cameras.

All of the existing works quoted above focused on the classification methods. However, we consider that the important elements for generic image classification/recognition are not only classification methods but also training images. Even if the classification method has high ability, the system cannot work for various images sufficiently in case that a diversity of learning images is not enough. A diversity of images real world scenes is extremely high in general, so that we have to gather as diverse images as possible to make system more practical. It is almost impossible to gather such various images of various kinds of real world scenes exhaustively by hand.

Then, we have proposed gathering visual knowledge for generic image classification/recognition of real world scenes from the World Wide Web[14]. We call this project "Web Image Mining for Generic Image Recognition". To say it concretely, our system utilizes images gathered automatically from the World Wide Web as training images for generic image classification instead of commercial image collections. 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 scenes, since Web images are taken by a large number of people for various kinds of purpose. It is expected that diverse training images enable the system to classify diverse real world images.

Assuming that we gather all the images in the whole world, a newly generated image must be very similar to some of the gathered images. This expectation is based on the heuristic that most of the scenes we see in our everyday life are similar to the scenes we have ever seen. In addition, people usually take pictures not at random but intentionally. They tend to place their targets on the center of pictures. In most case, the targets are parts of the real world scene such as objects and views. Although a diversity of scenes in the real world is huge, a diversity of images, which are taken as photos by the people, is less than that. Therefore, we expect that diverse training images must enable the system to classify diverse real world images.

In addition, the main targets of the conventional works on Web mining are numeric data and text data. However, there are a large number of multimedia data such as images, movies and sounds on the Web. The number of images on the Web especially is rapidly increasing recently because of the recent rapid spread of digital cameras. We think that use of multimedia data on the Web, namely visual knowledge on the Web, is promising and important for resolving real world image recognition/classification.

The processing in the system we are developing in our "Web Image Mining" project 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 image features and classes. The system is constructed by integrating three modules, which are an image-gathering module, an image-learning module, and an image classification module (Figure 1).

In this paper, we describe methods of image-gathering

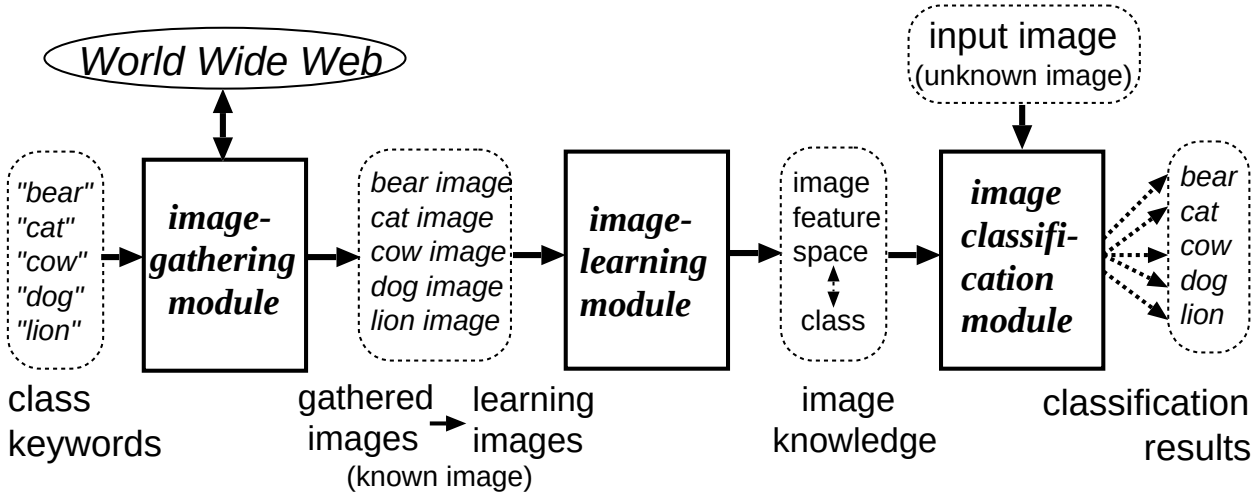

Figure 1: Web Image Mining system, which is constructed as an integrated system of an image-gathering module, an image-learning module and an image classification module.

from the World Wide Web and classification of unknown images using images gathered from the Web. They are slightly modified compared to [14]. The main difference is in the image-gathering module. We propose a method to gather more images with higher accuracy for the image gathering stage. Finally, we describe experimental results and conclusions.

2 Image Gathering

The image-gathering module gathers images from the Web related to class keywords. Note that we do not call this module the image "search" module but the image "gathering" module, since its objective is not to search for a few highly relevant images but to gather a large number of relevant images.

At present, some commercial image search engines on the Web such as Google Image Search, Ditto and AltaVista Image Search are available. Their preciseness of search results is, however, not good since they employs only keyword-based search. Then, some integrated search engines employing both keyword-based search and content-based image retrieval have been proposed. WebSeer [5], WebSEEk [11] and Image Rover [10] have been reported so far. These systems search for images based on the query keywords, and then a user selects query images from search results. After this selection by the user, the systems search for images that are similar to the query images based on image features. These three systems carry out their search in an interactive manner. Therefore, they are not suitable for "Web Image Mining", which requires gathering a large number of images.

Since an image on the Web is usually embedded in an HTML document that explains it, first the module exploits some existing commercial text-based Web search engines and gathers URLs (Universal Resource Locator) of HTML documents related to the class keywords. In the next step, using those gathered URLs, the module fetches HTML documents from the Web, analyzes them. If it is judged that images are related to keywords, the image files are fetched from the Web. Then, we remove irrelevant images from them based on their image features, and regard them as output images.

Removing irrelevant images is carried out by eliminating images which belong to relatively small clusters in the result of image-feature-based clustering. Images which are not eliminated are regarded as appropriate images to the class keywords, and we store them as output images. Our preference of larger clusters to smaller ones is based on the following heuristic observation: an image that has many similar images is usually more suitable to an image represented by keywords than one that has only a few similar images.

The processing of the image-gathering module consists of collection and selection stages.

In the collection stage, the system obtains URLs using some commercial web search engines, and by using those URLs, it gathers images from the web. The detail of the algorithm is as follows.

1. A user provides the system with two kinds of query keywords. One is a main keyword that best represents an image, and the other is an optional subsidiary keyword. For example, when we gather "apple" images, we use "apple" as a main keyword and "fruit" as a subsidiary keyword. Subsidiary keywords help to restrict the kind of gathered images. In this case, it prevents "apple computer" images from being gathered.
2. The system sends the main and subsidiary keywords as queries to the commercial search engines and obtains the URLs of the HTML documents related to the keywords.
3. It fetches the HTML documents indicated by the URLs.
4. It analyzes the HTML documents, and extracts the URLs of images embedded in the HTML documents with image-embedding tags ("`IMG SRC`" and "`A HREF`"). The system fetches files whose images satisfy one of the following conditions.

   Condition:

   - If the image is embedded by the "`SRC IMG`" tag, the

"ALT" field of the "SRC IMG" includes the keywords.

- If the image is linked by the "A HREF" tag directly, the words between the "A HREF" and the "/A" include the keywords.
- The name of the image file includes the keywords.

The reason why we set up the above conditions is that our preliminary experiments turned out that an ALT field, link words and a file name had high tendency to include keywords related to the image. In this paper we do not use the other clues such as a TITLE tag and frequency of keywords, because these clues have only low tendency to include keywords related to the image. Although the number of gathered images in this method is not large, we can compensate it with the re-gathering processing described later.

In the selection stage, the system remove irrelevant images from them based on their image features.

1. The system first makes image feature vectors for all the collected images. We use signatures and the Earth Mover Distance(EMD)[8] as image features and dissimilarity. Here, we compute one signature from one image. Next, all the distances (dissimilarity) between two images are calculated based on the EMD.
2. Based on the distance between images, they are grouped by the hierarchical cluster analysis method. Our system uses the farthest neighbor method (FN). In the beginning, each cluster has only one image, and the system repeats merging clusters until all distances between them are more than a certain threshold.
3. It throws away small clusters that have fewer images than a certain threshold value, regarding them as being irrelevant. All the images in the remaining clusters are regarded as output images.

In the image-gathering module, we can gather more images as we obtain more URLs of HTML documents. However, for one set of query keywords, the number of URLs obtained from Web search engines was limited because commercial search engines restrict the maximum number of URLs returned for one query. Thus, we introduce the query expansion method [9] for generating automatically new sets of query keywords for search engines.

The system extracts the top several words (only nouns, adjectives, and verbs) with high frequency except for initial query keywords from all HTML files with embedded output images of the initial image gathering, and regards them as subsidiary query keywords. It generates several sets of query keywords by adding each subsidiary words to the main keyword, and then obtains a large number of URLs for the query keyword sets. For carrying out the second image gathering, using obtained URLs, the system goes through the collection and selection stages again. This "query expansion and re-gathering" enables the number of images gathered from the Web to increase greatly.

3 Image Classification

First, in the learning stage, the image-learning module extracts image features from images gathered by the gathering module 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 corresponding to the class keywords by comparing image features.

We use Earth Mover Distance(EMD)[8] as image features and dissimilarity, and image-feature-based search which is a k-nearest neighbor variant as a classification method. It is the same as [14].

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 together 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 training images. Here, we do not set up "reject", and then all test images are classified into any class.

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. To obtain color signatures, first, we normalize the size of training images into 240×180, and divide them into 25 block regions. Next, we make a color signature for each of these 25 block regions by clustering color vectors of each pixel into color clusters by the k-means method.

In case of *region signatures*, a set of feature vectors of regions and their ratio of pixels represents a region signature. To obtain region signatures, we carry out region segmentation for images instead of dividing images into block segments after normalizing their size. Here, we employ a simple segmentation method based on the k-means clustering used in [13] and a sophisticated color segmentation method, JSEG [3].

To compute dissimilarity between two signatures, Earth Mover's Distance (EMD) has been proposed [8]. The EMD are found to be the most excellent distance on the average among distances commonly used in content-based image retrieval, as indicated by the prior work of Y.Rubner et al. [7].

In the classification stage, in case of the color signature, we sum up the minimum distances between an unknown input image and training images of each class for 25 all blocks, and classify it into the class whose total distance is the smallest. In case of the region signature, we employ the k-nearest neighbor (k-NN) method to classify an unknown input image into one of the class. The value of k is decided as 5 by the preliminary experiments.

Table 1: Four experiments.

no.	# of classes	# of images	precision (%)	test images #	test images source
1	20	3790	77.8	50	W+C†
2	20	11205	80.8	50	W+C†
3	20	34043	80.1	50	W+C†
4	50	17379	— ‡	LO	Web

†Web images + Corel images ‡not examined
CV: cross-validation, LO: leave-one-out

4 Experimental Results

We made four experiments from no.1 to no.4 as shown in Table 1. The experiment no.1, 2 and 3 are 20-class classification experiments, and the experiment no.4 is a 50-class classification experiment. In the experiment no.1, 2 and 3, we made the experiments using three different training image sets gathered from the Web independently. We used the query expansion technique for the experiments no.2 and no.3. In each experiments, we extracted the top 5 and 15 words from the initially-gathered HTML files.

In the experiment no.1, we gathered images from the Web for 20 kinds of class keywords shown in Table 2. By the image-gathering module about ten thousands URLs were fetched from three commercial text search engines, Google, InfoSeek, Goo Japan. The total number of gathered image was 3790, and the precision by subjective evaluation was 77.8%, 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 second and third columns of Table 2, we show the number of URLs of images gathered from the Web and their precision.

In the columns after the fourth of Table 2, we show the classification result using the gathered images from the Web as training images. Note that the precision of training images is not 100% unlike the conventional works on image classification. In the experiments no.1, 2 and 3, we used a special hand-made test image set for evaluation. We make a special test image set by selecting various kinds of 50 typical images for each class from Corel Image Gallery and Web images by hand. The table describes only results by color signatures in each class, since most of results by color signatures are superior to results by region signatures using k-means and JSEG. In the table, "region (1)" and "region (2)" mean region signature using the k-means clustering and region signature using the JSEG region segmentation method. In the tables, 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}, 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 37.3 as the F-measure value by color signatures.

In Table 2, we also show the results of the experiment no.2 and 3. Their differences from no.1 are only the number of training images. These results shows the F-measure rose as the number of training images increased. The results of "apple", "house", "car" and "Mt.Yari" especially were improved a lot. On the other hand, the classification rate of "Ichiro" remained low, since "Ichiro" images had much variation and no typical pattern. About 500 training images were not enough to classify them. These results indicate that the difficulty to classify images depends on the nature of the class greatly.

Table 3: Results of the experiment no. 4

method	exp. no.4 rec.	pre.	F
avg. by color	28.6	53.0	**37.1**
avg. by region (1)	24.5	27.9	**26.1**
avg. by region (2)	21.6	24.6	**23.0**

In the experiment no.4, we made a classification experiment for 50 class keywords, which were selected from words related to nature, artifacts and scene. We obtained 28.6, 53.0 and 37.1 as the recall, the precision and the F-measure, respectively, by color signatures (Table 3). This results are 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. It is hard to collect such various kinds of images as images used in the experiment no.4 by means of commercial image databases, and it has come to be possible by image-gathering from the World Wide Web.

5 Conclusions

In this paper, we described our ongoing project, "Web Image Mining for Generic Image Recognition". This project aims at generic image classification using images automatically-gathered from the Web as training images instead of hand-made image collections.

Although classification rate obtained in the experiments for generic real world images is not high and not sufficient for practical use, the experimental results suggest that generic image classification using visual knowledge on the World Wide Web is one of the promising ways for resolving real world image recognition/classification.

There are many issues to be solved for making this project more practical. How many classes does a generic classification system have to treat for practical use? How many training images are required for each class? What should we define as a "class"? Should we remove as many irrelevant images as possible from training image sets? What is the classification way to treat with training sets including some irrelevant samples well? Because of such issues, evaluation is the biggest problem for generic image classification/recognition. Experimental results sometimes depend on learning sets and training sets more greatly than classification algorithms. Therefore, we need comprehensive standard benchmark test sets like TREC.

Lately, Corel images are a de facto standard set for the evaluation of generic image classification systems. As we indicated in this paper, however, commercial photos such as Corel images are well-arranged, and most of ones in the

Table 2: Results of the experiment no. 1, 2 and 3.

	exp. no.1					exp. no.2					exp. no.3				
class	num.	pre.	rec.	pre.	F	num.	pre.	rec.	pre.	F	num.	pre.	rec.	pre.	F
apple	86	(86)	9.4	100.0	17.2	614	(86)	47.1	93.0	62.5	2561	(83)	55.3	67.1	60.6
bear	261	(47)	13.7	13.7	13.7	729	(76)	23.5	22.6	23.1	2533	(73)	29.4	17.4	21.9
mountain bike	82	(62)	1.5	100.0	3.0	575	(78)	6.2	66.7	11.3	2386	(88)	23.1	19.0	20.8
Lake Biwa [a)]	272	(38)	32.4	30.7	31.5	679	(49)	22.5	47.1	30.5	2191	(62)	67.6	30.8	42.3
car	41	(76)	0.0	0.0	0.0	694	(92)	38.9	26.9	31.8	1836	(92)	61.1	30.0	40.2
cat	267	(82)	23.5	21.4	22.4	611	(92)	15.7	28.6	20.3	2532	(84)	49.0	19.1	27.5
entrance [b)]	440	(96)	89.7	24.1	38.0	648	(94)	46.6	49.1	47.8	1934	(92)	70.7	46.1	55.8
house	33	(79)	1.6	100.0	3.2	306	(82)	0.0	0.0	0.0	2575	(77)	19.7	25.0	22.0
Ichiro [c)]	75	(80)	0.0	0.0	0.0	392	(68)	3.5	66.7	6.7	481	(61)	1.8	100.0	3.4
Ferris wheel	81	(89)	9.0	77.8	16.1	395	(74)	19.2	83.3	31.2	731	(83)	12.8	100.0	22.7
Kinkaku Temple [d)]	222	(89)	73.8	68.2	70.9	352	(86)	23.0	100.0	37.3	700	(75)	36.1	95.7	52.4
lion	70	(86)	23.5	75.0	35.8	481	(75)	23.5	41.4	30.0	1738	(92)	27.5	93.3	42.4
Moai	104	(78)	29.4	93.8	44.8	329	(93)	49.0	89.3	63.3	549	(87)	43.1	91.7	58.7
note-size PC	83	(55)	10.8	77.8	18.9	499	(58)	15.4	83.3	26.0	852	(59)	23.1	71.4	34.9
Shinkansen train [e)]	98	(66)	11.1	75.0	19.4	548	(77)	16.7	60.0	26.1	760	(65)	11.1	85.7	19.7
park	328	(83)	62.9	28.9	39.6	762	(89)	95.2	17.0	28.8	2193	(88)	58.1	26.1	36.0
penguin	195	(91)	13.0	43.8	20.0	441	(78)	1.9	50.0	3.6	1791	(82)	14.8	88.9	25.4
noodle [f)]	365	(90)	78.6	46.6	58.5	786	(88)	40.0	48.3	43.8	2212	(96)	42.9	62.5	50.8
wedding	125	(88)	7.0	40.0	11.9	645	(89)	31.6	37.5	34.3	2062	(78)	28.1	37.2	32.0
Mt.Yari [g)]	562	(94)	94.5	13.0	22.9	719	(93)	92.7	15.1	26.0	1426	(85)	72.7	38.8	50.6
total avg. by color total	3790	(78)	29.3	51.5	**37.3**	11205	(81)	30.6	51.3	**38.3**	34043	(80)	37.4	57.3	**45.2**
avg. by region (1)			29.7	34.0	**31.7**			33.1	36.2	**34.6**			35.3	40.1	**37.6**
avg. by region (2)			28.6	31.9	**30.2**			30.8	32.0	**31.4**			31.2	31.5	**31.4**

a) the biggest lake in Japan b) a school entrance ceremony c) the name of a famous baseball player d) a famous temple in Japan
e) Japanese bullet train f) Chinese noodle g) a famous mountain in Japan

same category are similar to each other unlike Web images which are diverse real world images. We think that a generic image classification system should be able to treat diverse real world images like Web images. For this, the system requires a large amount of visual knowledge for many classes. Since the World Wide Web has them, visual knowledge on the Web will enables us to realize a generic image recognition system.

References

[1] K. Barnard and D. A. Forsyth. Learning the semantics of words and pictures. In *Proc. of IEEE International Conference on Computer Vision*, volume II, pages 408–415, 2001.

[2] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Recognition of images in large databases using a learning framework. Technical Report 07-939, UC Berkeley CS Tech Report, 1997.

[3] 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.

[4] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicons for a fixed image vocabulary. In *Proc. of European Conference on Computer Vision*, 2002.

[5] C. Framkel, M. J. Swain, and V. Athitsos. WebSeer: An image search engine for the World Wide Web. Technical Report TR-96-14, University of Chicago, 1996.

[6] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proc. of First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[7] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding*, 84(1):25–43, 2001.

[8] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[9] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[10] S. Sclaroff, M. LaCascia, S. Sethi, and L. Taycher. Unifying textual and visual cues for content-based image retrieval on the World Wide Web. *Computer Vision and Image Understanding*, 75(1/2):86–98, 1999.

[11] J. R. Smith and S. F. Chang. Visually searching the Web for content. *IEEE Multimedia*, 4(3):12–20, 1997.

[12] J. Z. Wang and J. Li. Learning-based linguistic indexing of pictures with 2-D MHMMs. In *Proc. of ACM International Conference Multimedia*, pages 436–445, 2002.

[13] J. Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.

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