Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing (CIISP 2007)

A Novel Approach to Collect Training Images from WWW for Image Thesaurus Building

Joohyoun Park and Jongho Nang

Dept. of Computer Science and Engineering
Sogang University
Seoul, Korea
{parkjh, jhnang}@sogang.ac.kr

***Abstract*—This paper introduces a novel approach to change gathered images from WWW into training images to build an image thesaurus. The requirements for being training images are a large number of images and with highly relevant to a given concept. To fulfill these requirements, a system should be able to collect a large number of relevant images to a given concept from WWW by *the proposed criterion of relevance to the concept for each image*. Then, the irrelevant images would be filtered out by *the modified hierarchical clustering method based on the weighted combination of 5 MPEG-7 visual descriptors*[9] and *the proposed criterion of relevance to the concept for each cluster*. Upon experimental results, the precision of the set of images generated by the proposed method is about 18% higher than that of the set of images generated by other methods[1][2].**

***Index Terms*—Auto Image Annotation, Content Based Image Retrieval**

I. Introduction

Recently, image thesaurus have been used as one of the feasible solutions to solve 'Semantic Gap[3]' problem in various CBIR (Content Based Image Retrieval) systems. The image thesaurus is the set of "visual keyword[5]", which represents the relationship between visual information (such as color, texture, shape, and edge) and a concept (textual information). Since many researches[12][13][14] regard the problem of building visual keywords as a supervised or an unsupervised learning problem, a large number of labeled images are essential for learning various concepts. Unfortunately, it is not easy to prepare a large number of labeled images for building visual keywords because it requires huge amount of time to put annotation manually. Commercial image collections such as Corel set may also have the lack of diversity to classify real world images because the images in the collection may be well arranged with good quality. For this reason, *images on WWW* could be a good solution to cope with these difficulties.

To use web images as training images to learn the relationship between low-level visual features and high level concepts, *a large number of web images have to be collected and annotated automatically with high precision*. There were some researches for automatic web image annotations[15][16][17] by analyzing HTML document. However, this approach would generate many irrelevant annotations as well as relevant ones because of the lack of measure which evaluates the degree of relevance between the surrounding texts and the images. [1] uses the labeled images which are annotated by this approach as training images directly even though there are many noise images. A large number of mismatched training samples would make the performance of image thesaurus worse. To improve the precision of these set of training images, [2] clusters gathered images into certain groups of images based on LUV color histogram and filter out images from relatively small clusters. However, it could not effectively eliminate the irrelevant images because of the use of single feature and also the lack of measure which evaluates the degree of relevance of clusters.

This paper proposes a new method to changing gathered images from WWW into training images to build an image thesaurus. The proposed method consists of two processes. First process is generating the set of candidate images for a given concept, which are collected from WWW by analyzing HTML document. For criterion of relevance to a concept for each image, the degree of relevance to the concept is evaluated according to the weighting factors such as the visual distance between an embedded image and the concept, HTML tags, and surrounding texts. Second process is filtering out irrelevant images from the set of candidate images by clustering based on five MPEG-7 visual descriptors [9]. To find the proper descriptors which have the key of classifying images for the concept, *CH Index*[4], the method of evaluating the partitioning obtained by clustering, is used. Based on the descriptors with high *CH Index*, images in the candidate set are classified into certain groups of images by similar visual characteristics. Finally, irrelevant clusters, which are determined by the size of cluster and the relevance of images in cluster, will be discarded. Upon experimental results, the precision of the set of images generated by the proposed method is about 18% higher than the set of images generated by [1] and [2].

II. Generating the Set of Candidate Images

In this section, we will present the way to collect related images to a given concept from WWW by analyzing HTML documents. the basic idea is similar to [6].

This work was supported by the Brain Korea 21 Project in 2006.

TABLE I
THE WEIGHTS BASED ON HTML TAGS

HTML Tags	Weights
<ALT>	1.0
<IMG>	0.845
<TITLE>	0.602
<Hx> or <ALT> or <I>	0.477
URL	0.477
No Tag	0

A. The Degree of Relevance to a Concept for an Image

To find the high relevant images to a given concept from WWW, the degree of relevance to the concept for an image has to be calculated with considerations for the following conditions.

First, considering the distance from an image to the concept[1], the words closer to the image could be more relevant. According to VIPS(VIsion based Page Segmentation)[10], a web page can be divided into several VB(Visual Block)s with high DoC(Degree of Coherence) which implies that the words in the VB embedding an image are more relevant to the image than words in other VBs. Second, words with specific HTML tags may be more relevant. For example, the words appearing with *src*, *alt* fields of the *img* tag, and *headers* may have higher importance than other words. **Table 1** shows the weights of relevance to an image based on HTML tags. Finally, the frequency of word (term frequency) has to be considered. As the number of times of showing in the HTML document grows, the importance of the word increases as well. Also, the synonyms of a given concept, which can be judged by WordNet[11], are regarded as the same word with the concept. For example, the appearance of "doggie" or "puppy" also increases the frequency of "dog" when the given concept is "dog".

Assuming that the robot collects relevant images with a given concept c from WWW. For a collected image o, the degree of relevance to c, w_{oc}, is calculated as follows;

$$w_{oc} = \begin{cases} \max(w_c, \min(\log(tf_c + 1), 1)) \\ \qquad\qquad \text{,if } c \text{ is in same VB with } o \\ 0 \qquad\qquad \text{,if } c \text{ is in different VB with } o \end{cases} \quad (1)$$

Note that tf_c is the frequency of c (includes the synonyms of c) in the VB embedding the image o and w_c is the weight based on HTML tags (see **Table 1**). Based on [6], the *log entropy* scheme and the weights were given based on HTML tags were given accordant with this scheme.

B. Generating The Set of Candidate Images for Concept

The set of candidate images for a concept c ($\Omega_c = \{o_i | w_{o_i c} > T, 1 \le i \le n\}$, where o_i is i^{th} image.) consists of the images whose the degree of relevance to the concept is higher than the given threshold T $(0 \le T \le 1)$ from n gathered images from WWW. T determines the characteristics of the set of candidate images such as a set of the small number of highly relevant images or a set of the large number of images with many noisy ones. That is, T is close to 1 if the objective of image collecting is searching for highly relevant images regardless of the size of set, and is close to 0 otherwise.

[1] In HTML document, a concept is represented as an word in surrounding texts

III. GENERATING THE SET OF TRAINING IMAGES

In this section, we will show the way to pick actual relevant images to a concept among the set of candidate images without any prior knowledge from the concept. This process would be regarded as unsupervised classification problem into two categories such as the group of relevant images and the group of irrelevant ones. The precision of training images could be improved dramatically, if the images in the relevant groups were picked as the training images. Assuming that images associated a concept have some peculiar visual characteristics, the group of actual relevant images can be achieved by clustering based its the visual features. Of course, in order that this idea works effectively, the two issues such that which features are used to classify images and which clusters are more relevant ones than others have to be considered carefully. the detail of the former and the latter will be discussed in section 3.1 and 3.2 respectively.

A. Clustering based on Multiple Visual Features

5 visual descriptors defined in MPEG-7[9] such as *Dominant Color (DC)*, *Color Layout (CL)*, *Color Structure (CS)*, *Edge Histogram (EH)*, and *Homogeneous Texture (HT)* were used as visual features. Even though these descriptors are good representations for images which have been tested on large data sets with a good performance during the process of standardization, not all the descriptors are the best descriptors for classification of images for each concept. For example, in case of *Dominant Color*, it could be the best descriptor for the concept "tiger" because of the peculiar color of "tiger", but it may be not for the concept "cat" because of its diversity of color of "cat". Therefore, in terms of the classification capability, *Homogeneous Texture* would be better than *Dominant Color* for this concept. Consequently, the best descriptors with high classification capability should be picked to classify images effectively for each concept.

According to [7], good descriptors should be able to generate descriptions with high variance, evenly balanced cluster structure and high discriminance capability to distinguish different content. Assuming that the best descriptors with high classification capability is also defined by [7], it could be found by evaluating the results achieved by clustering based on each descriptor. In this system, *CH Index*[4], the method of evaluating the partitioning obtained by clustering, is used to select a good descriptor for each concept. Since this index is a function of the ratio of the between cluster separation to within cluster scatter, high index value implies high classification ability. It implies

that a descriptor with high index value would be a good candidate and it could be proved by the preliminary results of experiment in most of the cases. However, the good descriptors judged by *CH Index* matches with the best descriptors picked manually with 70% concepts while 30% concepts do not match with it. Therefore, the clustering results for 30% concepts would be very poor. Instead of using a good feature, The use of combined multiple features with dynamically updated weights specified by *CH Index* could solve this problem.

Although a good descriptor may not be the best for every case, it could give good indication on the importance of the descriptors during the classification process. The distance function $d(i,j)$ between two image i and j can be written as follows;

$$d(i,j) = \sum_x \frac{ch_x}{\lambda} \cdot GausNorm(d_x(i,j)) \\ \text{,where } \lambda = \sum_x ch_x \text{ , } x = \{DC, CL, CS, EH, HT\} \quad (2)$$

Note that d_x and ch_x are the distance function defined in the MPEG-7 visual part of eXperience Model (XM)[8] and *CH Index* value for a descriptor x respectively. $GauseNorm$ means Gaussian Normalization[18] which normalized the distance of each descriptor within [0,1].

Fig.1 shows an algorithm to classify the set of candidate images which are collected by analyzing HTML document into certain image groups with similar visual features for a concept. The results of hierarchical clustering for each descriptor would be evaluated by *CH Index*. In **HClustering**, *CH Index* is evaluated while the number of clusters varies from max_k to min_k. Then, the maximum value of *CH Index* is picked as the representative index value for the descriptor x because the number of clusters k' that maximizes the value of *CH Index* is taken as the optimal number of clusters. Finally, images in Ω_c are classified into k' image groups $C_i (1 \le i \le k')$ based on the distance function d for 5 visual descriptors with weights specified by these index values (see **Eq.2**).

```
//c : the given concept
//Ω_c : the set of n candidate images for concept c
//o_i : the i^th image in the set
//C_i : the i^th cluster
//min_k : the minimum number of clusters
//max_k : the maximum number of clusters
//CH[k] : CH index value when the number of clusters

Clusters HClustering(metric[in] d, real[out] ch)
begin HClustering
for all o_i ∈ Ω_c such that 1 ≤ i ≤ n do
  C_i = {o_i}; //initially make n clusters
end for
for k = n − 1 to min_k do
  find nearest clusters, say, C_l and C_m using d;
  merge C_l and C_m;
  if k ≤ max_k then
    CH[k] = compute CH Index for k clusters;
  end if
end for
k' = arg_k max{CH[k]};
ch = CH[k'];
return Clusters when the number of clusters is k';
end HClustering

Clusters ClusterCandidateSet()
begin ClusterCandidateSet
for x=DC, CL, CS, EH, HT do
  HClustering(d_x, ch_x);
end for
//make the metric d for multi descriptors
d = Σ_x (ch_x/λ) · d_x, where λ = Σ_x ch_x;
Φ_c = HClustering(d, ch);
return Φ_c
end ClusterCandidateSet
```

Fig. 1. Algorithm to Cluster the Set of Candidate Images based on Multiple Visual Descriptors

B. Selecting Relevant Clusters

Let Φ ($\Phi = \{C_i, 1 \le i \le k'\}$, where C_i is i^th cluster) be the set of clusters achieved by the procedure **ClusterCandidateSet**. To cut off the irrelevant clusters from the final set A_c for a concept c, the degree of relevance, $S_c(C_i)$, has to be evaluated for each cluster C_i. Since it is difficult to know the patterns of visual features associated the concept at this time, $S_c(C_i)$ has to be evaluated based on the degree of relevance of images (see **Eq.1**) in C_i.

Fig.2 shows the precision as a function of the degree of relevance to each concept for 6,400 images based on the result of preliminary experiments. As shown in this graph, the precision increases proportionally as the relevance to concept of images increases. According to these results, the probability of being a relevant image to a concept can be calculated when the degree of relevance is varies. The $E_c(C_i)$, which is the expectation value of being a relevant image could be picked up randomly from cluster C_i for a concept c, and estimated by the average of these probabilities for all images in the cluster.

In order to use the expectation value as the degree of relevance for each cluster, the size of cluster has to be considered. In the following example, the set of candidate images Ω_c is classified into two clusters C_1 and C_2, which consist of two sets of images with the degree of relevance 1 and 10 images with 0.8 respectively. The system will discard C_2 because its the expectation value is lower than C_1. This case is not desirable but happening. To avoid this undesirable result, the degree of relevance for a cluster C_i, $S_c(C_i)$, should be defined as follows;

$$S_c(C_i) = \frac{\log n_{C_i}}{\log n_{C_i} + 1} \cdot E_c(C_i) \quad (3)$$

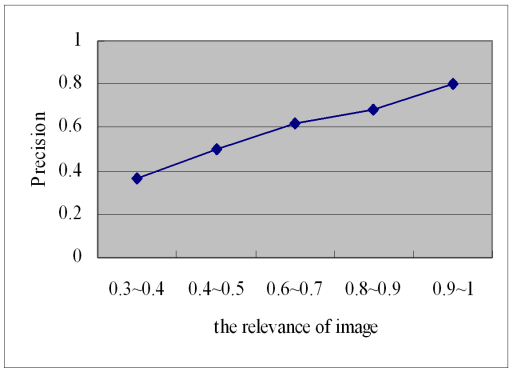


Fig. 2. The precision as a function of the degree of relevance to each concept

Note that n_{C_i} is the number of images in C_i. To avoid highly sensitive changes to the number of images, log scheme was chosen. The term $\log n_{C_i} + 1$ is to normalize $S_c(C_i)$ within [0,1].

Fig.3 shows an algorithm to build the set of training images based on the degree of relevance to a concept. For a cluster C_i in Φ_c, All images are stored as training images when $S_c(C_i)$ is larger than the given threshold T_1. Every other images in $S_c(C_i)$ has to be tested to be a training image individually.

```
//c : the given concept
//A_c : the final set of training images for concept c
//Φ_c : the set of clusters
//o_j : the j^th image in the set
//C_i : the i^th cluster
//S_c(C_i) : the degree of relevance to the concept for C_i

Set of Images MakeTrainingImages(Clusters Φ_c)
begin MakeTrainingImages
A_c = {};
for all C_i ∈ Φ_c such that 1 ≤ i ≤ k do
  Calculate S_c(C_i);
  if S_c(C_i) ≥ T_1 then
    A_c = A_c ∪ C_i
  else
    for all o_j ∈ C_i do
      if w_{o_j c} ≥ T_2 then
        A_c = A_c ∪ o_j
      end if
    end for
  end if
end for
return A_c
end MakeTrainingImages
```

Fig. 3. Algorithm to Build the Set of Training Images

IV. Experiments

We gathered images on WWW for 10 kinds of concept such as "tiger", "tree", "bird", "reptile", "dog", "snake", "cat", "flower", "insect", and "mountain". To generate the sets of candidate images for these concepts, image robot has to visit enormous number of web pages by BFS(Breadth First Search) traversal method from 6 seed sites and collected 6,400 images as shown in **Table 2**.

TABLE II
Gathered images from 6 seed sites for the experiments

Seed site	# of images
http://www.junglewalk.com	647
http://nationalzoo.si.edu	1192
http://www.freefoto.com	1228
http://www.hickerphoto.com	1379
http://www.amusetoi.com	1586
http://www.indianwildlifeportal.com	708

Fig.4 shows the results of gathered candidate images for 10 different kinds of concepts from WWW. Every concepts except for the "tree" and "flowers", when the threshold T is 0.4 compared to $T = 0.3$, the precision increases prominently as shown in **Fig.4-(a)**. On the other hand, the recall[2] decreases dramatically as shown in **Fig.4-(b)**. The result indicate that discarding a large number of relevant images to a concept helps improving the precision because not a few images whose the degree of relevance the concept is within [0.3, 0.4)[3] are relevant actually. For example, in case of "tree" and "flower", the recall decreased slightly because these concepts are not main objects in images even though the words "tree" and "flower" appear in the surrounding texts such as "bees in flowers", "lion under trees", ... etc.

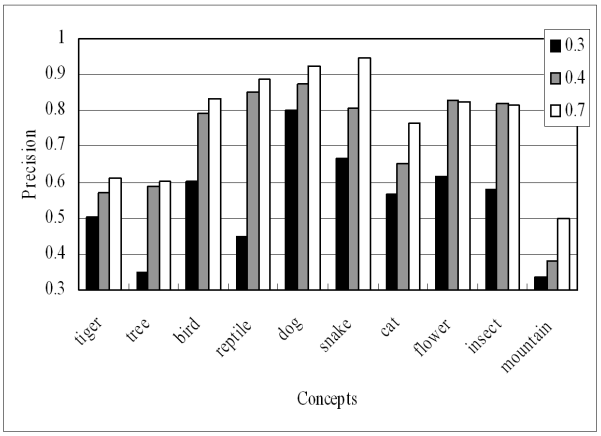
Now, let us experimentally compare the precision of the image set generated by the proposed method, keiji's method [2], and analyzing HTML document(**candidate set**). For all experiments, 0.1 and 0.8 were given to T_1 to cut off clusters with low relevance and T_2 to save images with high relevance in dropped clusters respectively. **Fig.5** shows the precision of the three methods when $T = 0.3$ (**Fig.5**-(a)) and $T = 0.7$ (**Fig.5**-(b)). As shown in these figures, the precisions of the proposed method are much higher than other methods for all concepts regardless of T. On the other hand, the precisions of Keiji's method are lower than those of candidate set for some concepts which indicate every small clusters can not be considered as an irrelevant ones for all concepts.

V. Conclusion

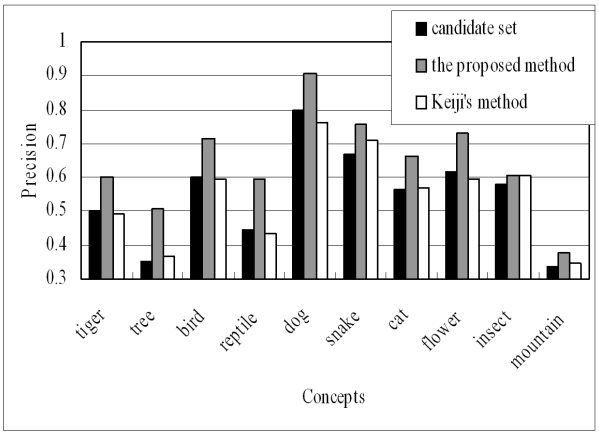
We presented a mechanism of changing gathered images from WWW into training images to build an image

[2] Since the recall is evaluated by the ratio of the relevant images to the ground truth which are selected manually when $T = 0.3$, the recall is 1 for all concepts when $T = 0.3$.

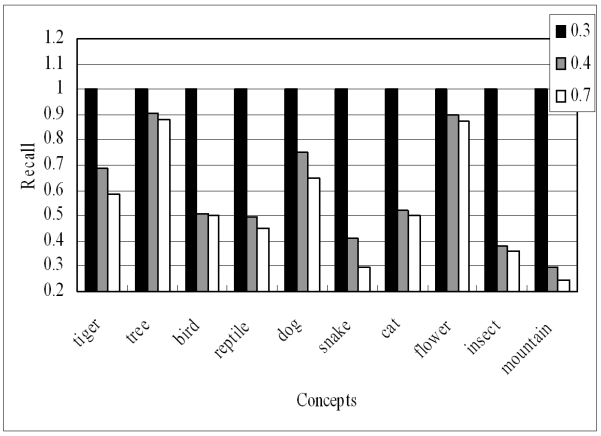
[3] The case that the degree of relevance to a concept is within [0.3, 0.4) is that the concept has to be occurred once in surrounding texts.



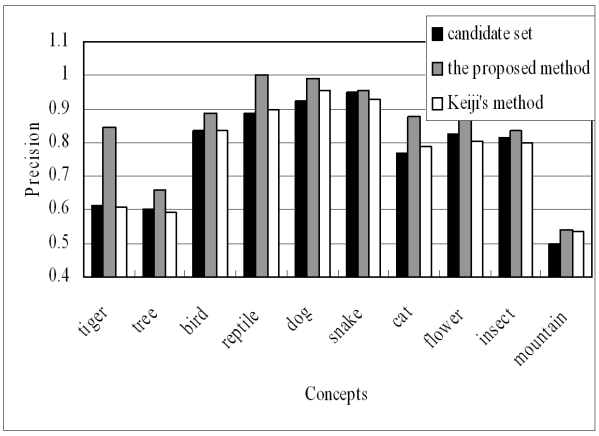
(a) The precision



(a) when $T = 0.3$



(b) The recall



(b) when $T = 0.7$

Fig. 4. The results of gathered candidate images for 10 kinds of concepts on WWW when T is 0.3, 0.4, and 0.7

Fig. 5. Experimental comparison of precision with the candidate set, the proposed method, Keiji's method [2]

thesaurus. To improve the quality of gathered images on WWW, *the criterions of relevance to a given concept for each image*, *the modified clustering method based on the weighted combination of five MPEG-7 visual descriptors*, and *the criterion of relevance to the concept for each cluster* were proposed as well. Based on the proposed methods, the system could generate a large number of training images with high precision and these images could be used to build an image thesaurus effectively.

References

[1] X. Wang, W. Ma, and X. Li, "Data Driven Approach for Bridging the Cognitive Gap in Image Retrieval", *In proceedings of ICME 2004*, pp.2231-2234, 2004.

[2] K. Yanai, "Generic Image Classification Using Visual Knowledge on the Web", *In Proceedings of ACM MM 2003*, pp.167-176, 2003.

[3] R. Yates and B. Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[4] U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.24, No.12, pp.1650-1654, 2002.

[5] J. Wu, M. Kankanhalli, J. Lim, and D. Hong, *Perspective on Content Based Multimedia Systems*, Kluwer Academic Publishers, 2000.

[6] M. Cascia, S. Sethi, and S. Sclaroff, "Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web", *In Proceedings of IEEE WorkShop on Content-based Access of Image and Video Libraries*, pp.24-28, 1998.

[7] H. Eidenberger, "Statistical Analysis of Content-based MPEG-7 Descriptors for Image Retrieval", *ACM Multimedia Systems Journal*, Vol.10, No.2, 2004.

[8] ISO/IEC JTC1/SC29/WG11, *MPEG-7 Visual part of eXperience Model Version 11.0*, 2001.

[9] ISO/IEC JTC1/SC29/WG11, *Information Technology Multimedia Content Description Interface-Part3: Visual*, 2001.

[10] D. Cai, S. Yu, J. Wen, and W. Ma, "VIPS: A Vision-based Page Segmentation Algorithm", *MSR-TR-2003-79*, 2003.

[11] C. Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, pp.265-283, 1998.

[12] S. Rui, W. Jin, and T. Shua, "A Novel Approach to Auto Image Annotation Based on Pair-wise Constrained Clustering and Semi-naive Bayesian Model", *In Proceedings of IEEE 11th International Conference on Multimedia Modeling*, pp.322-327, 2005.

[13] L. Wang, L. Liu, and L. Khan, "Automatic Image Annotation and Retrieval Using Subspace Clustering Algorithm", *In Proceedings of ACM MMDB 04*, pp.100-108, 2004.

[14] Y. Mori, H. Takahashi, and R. Oka, "Image-To-Word Transformation based on Dividing and Vector Quantizing Images with

Words", *In Proceedings of International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[15] J. Smith and S. Chang, "WebSeek: An Image and Video Search Engine for the World Wide Web", *IS&T/SPIE Proceedings of Storage and Retrieval for Image and Video Database*, 1997.

[16] C. Frankel, M. Swain, and V. Athitsos, "WebSeer: An Image Search Engine for the World Wide Web", *Technical Report 96-14*, University of Chicago Computer Science Department, 1996.

[17] N. Rowe and B. Frew, "Automatic Caption Localization for Photographs on World Wide Web Pages", *Information Processing and Management*, Vol.34, No.1, 1997.

[18] Y. Rui, T. Huang, and S. Mehrota, "Content based Image Retrieval with Relevance Feedback in MARS", *In Proceedings of International Conference on Image Processing*, pp.815-818, 1997.