Web Image Search by Automatic Image Annotation and Translation

Jin Hou[*1,2], Dengsheng Zhang[3], Zeng Chen[1], Lixing Jiang[1], Huazhong Zhang[1], Xue Qin[1]
[1]School of Information Science and Technology, Southwest Jiaotong University,
Chengdu, Sichuan 610031, P.R. China
jhou@swjtu.edu.cn
[2]State Key Laboratory for Novel Software Technology, Nanjing University,
Nanjing, Jiangsu 210093, P.R. China
[3]Gippsland School of IT, Monash University, Churchill, VIC 3842, Australia
dengsheng.zhang@infotech.monash.edu.au

Abstract—There has been a growing interest in implementing online Web image search engine in the semantic level. However, most existing practical systems including popular commercial Web image search engines like Google and Yahoo! are either text-based or a simple hybrid of texts and visual features. This paper proposes a novel Web Image Search by Automatic Image Annotation and Translation (WISAIAT) system by using automatic image annotation and translation. We develop a technology which learns semantic image concepts from image contents and translates unstructured images into textual documents, so that images are indexed and retrieved in the same way as textual documents. Existing database management systems can be used to effectively manage image contents and image search can act as efficient as text search by translating images to textual documents through large scale machine learning. Experiments in both the Corel dataset and real Web dataset are performed to validate our system and the results are promising. This system suggests a new combination of texts and visual features to achieve a semantic Web image search and expected to become a reranking system to the existing Web image search result available online via the Internet.

Keywords- semantic image search; automatic image annotation; image translation; Web image reranking; decision tree.

I. INTRODUCTION

With the number of digital images in the WWW increasing explosively, efficient image search in large-scale datasets has attracted great interest from both academia and industry. However, image retrieval is currently far less efficient than text retrieval because images are unstructured and much more difficult to process than texts. The approaches of retrieving and ranking images from large-scale datasets can be largely divided into the following three categories:

- Text-based approaches: the search engine returns corresponding images by processing the associated textual information, such as file name, surrounding text, URL, etc., according to keywords input by users [1]-[3]. Most of popular commercial Web image search engines like Google and Yahoo! adopt this method. While text-based search techniques have been verified to perform well in textual documents, they often result in mismatch when applied to the image search. The reason is that metadata can not represent the semantic content of images. For example, a search by the keyword "tiger" nets a large number of images of a golf player Tiger Woods and the animal tigers in the meantime.
- Content-based approaches: the search engine extracts semantic information from image content features, such as color, shape, texture, spatial location of objects in images, etc [4]-[8]. The extracted visual information is natural and objective, but completely ignores the role of human knowledge in the interpretation process. As the result, a red flower may be regarded as the same as a rising sun, and a fish the same as an airplane etc.
- Hybrid approaches: recent research combines both the visual content of images and the textual information obtained from the Web for the WWW image retrieval [9]-[11]. Such methods exploit the usage of the visual information for refining the initial text-based search result. Especially, through user's relevance feedback, i.e., the submission of desired images or visual content-based queries, the reranking for image search results can achieve a significant performance improvement.

In this paper, we propose a novel Web Image Search by Automatic Image Annotation and Translation (WISAIAT) system by translating images to texts rather than simply combining visual features and metadata. We employ the state of the art of machine learning technology to learn semantic image concepts from image contents so as to make images be indexed and retrieved like texts. An automatic annotation method by hybriding decision tree (DT) and support vector machine (SVM) is proposed and a novel inverted file is used to rank the search result. Experiments of both word search and image search in a Corel dataset and a Yahoo! dataset are performed. The preliminary result is satisfied and promising.

The rest of this paper is organized as follows. Section II introduces the prototype architecture and the main methodologies. Section III demonstrates the effectiveness and evaluates the performance of our

[*]Corresponding author, Email: jhou@swjtu.edu.cn, Tel: +86-28-87601742

system with experiments. Finally, section IV concludes the paper.

II. Methodology

This section addresses the key methodologies applied in WISAIAT, including image translation architecture, automatic annotation technology, and image indexing approach using an inverted file.

A. *Image Translation*

As mentioned above, images are unstructured and far more complicated to index and retrieve than texts. In this research, we propose a novel image translation architecture illustrated in Fig.1. We adopt a purely object-oriented approach , i.e., images are translated into individual objects in a dictionary and then indexed and retrieved in the same way as texts, which is our proposed method most prominent difference from the existing systems. The system prototype consists of six major modules: region representation, dictionary building, automatic annotation, invert file indexing, similarity calculation and user interface. WISAIAT employs two types of datasets as training datasets: (1) a set of 5100 images collected from Corel photo gallery, and (2) a set of 5260 images collected from Yahoo! image search. Images are segmented to object regions by the segmentation technology first. Our previous experiment shows that curvelet transform gives accurate edge information, especially at the highest scale where object contour is produced. The segmentation result is improved by combining with the edge information generated by curvelet transform, and the segmented regions are closer to semantic objects in images. Then visual features like color, texture and shape in each region are extracted and discretized as the input of automatic annotation.

B. *Automatic Annotation*

Most of current approaches use complex Bayesian models to infer the correlations or joint probabilities between images and annotations [12-14]. The problem with those models, however, is that they are arbitrarily formulated and application dependent. There are no consistent modeling of prior and conditional probabilities in literature. Consequently, a Bayesian model formulated for one image database may not be useful when it is applied to another image database. This paper presents an automatic annotation approach using DT plus SVM. DT is one of the most popular classification algorithms in data mining and machine learning. DT has semantic interpretation power which is a natural simulation of human learning, and does not make a priori assumption about the application problems. DT classifies instances by traversing from root node to leaf node to build a tree. Also, we perform split criteria and pruning technology in the tree to reduce the size of the tree. Fig.2 gives an example of the DT, where “color” and “texture” are the vectors of the attributes, and “mountain”, “horse”, “rock”, etc. represent the instances classified by the rule of the tree. In addition, we employ SVM technology before building the DT to improve the effectiveness of the classification. Our experiment shows that the classification accuracy and annotation accuracy can be improved around 44% and 8% respectively.

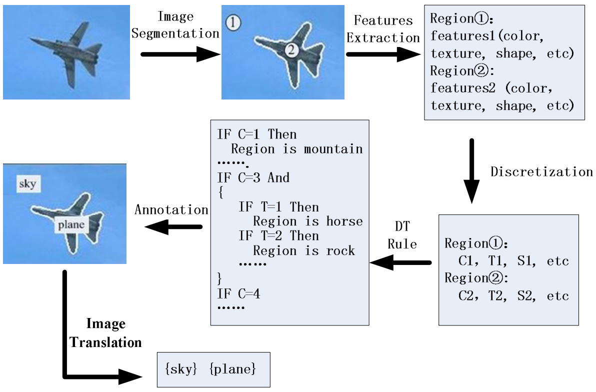

Figure 1. Image translation architecture.

The annotation using DT consists of two stages. The first stage is the training stage when the DT is trained by annotating the images in the training dataset. Once the DT is trained, it is used to annotate unknown images. Extensive investigation and rigorous validation tests are conducted in the training stage to achieve high annotation accuracy. Fig.3 shows an automatic annotation result for an image at the training stage.

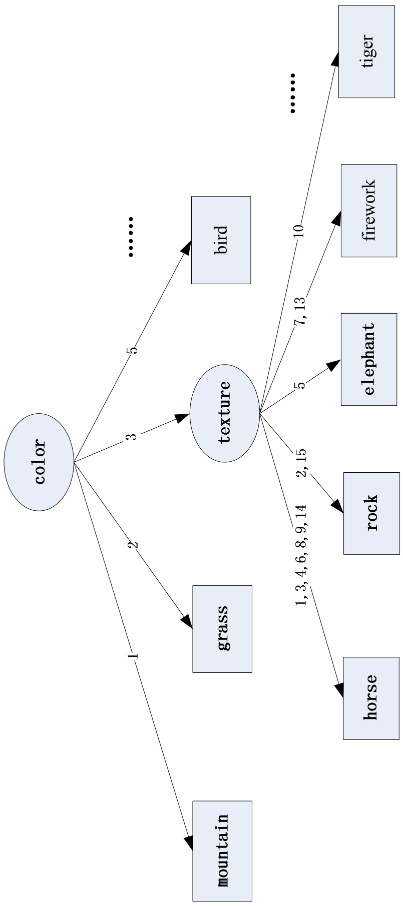

Figure 2. Decision tree.

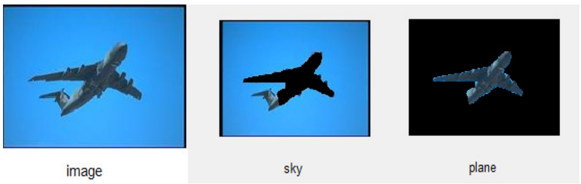

Figure 3. Image automatic annotation result.

C. *Image Indexing Using Inverted File*

Contrary to the conventional indexing method in the form of {image × terms}, a region-based indexing approach using an inverted file in the form of {term× images} is proposed, as illustrated in table I, where *df$_j$* represents the image frequency, i.e., the number of the images containing the *term$_j$*, tf_j^i stands for the term frequency, i.e., the number of times *term$_j$* appears in the i^{th} image, and the region information can be calculated by the equation (1).

$$\mathrm{regInfo}_j^i = \{(a_1^i, p_1^i, r_1^i), (a_2^i, p_2^i, r_2^i), \ldots, (a_{tf_j^i}^i, p_{tf_j^i}^i, r_{tf_j^i}^i)\}$$

$$= \sum_{k=1}^{tf_j^i} (area\_weight \times pos\_weight_k^j \times rel\_weight_k^j) \qquad (1)$$

According to the equation (1), we know that the region information is decided by the general weight of image area, position and relationship.

Then, the images are listed in descending order of similarity for each term. The similarity computation formula is given as

$$Similarity = \frac{\sum_{1\le j\le M} tw_j^i \times q_j}{\sqrt{\sum_{1\le j\le M} (tw_j^i)^2 \times \sum_{1\le j\le M} q_j}},$$

$$tw_j^i = regInfo_j^i \times idf_j, \quad idf_j = \log(\frac{N}{df_j}),$$

$$q_j = \begin{cases} 1, & \text{if } term_j \text{ appears in the query text} \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

where N and M are the total numbers of images and terms respectively (M<<N). Finally, images are indexed and retrieved based on the similarity values. The bigger the value is, the easier the image is retrieved. More details about this algorithm can be found in our previously published paper [15].

TABLE I. THE INVERTED FILE FOR IMAGE INDEX

Terms	**Images** ($df_j, tf_j^i, regInfo_j^i, tw_j^i$)
$term_1$	$\langle im_1, Simi_1^1 \rangle, \langle im_2, Simi_1^2 \rangle, \cdots, \langle im_{df_1}, Simi_1^{df_1} \rangle$
$\vdots$	$\vdots$
$term_j$	$\langle im_1, Simi_j^1 \rangle, \langle im_2, Simi_j^2 \rangle, \cdots, \langle im_{df_j}, Simi_j^{df_j} \rangle$
$\vdots$	$\vdots$
$term_M$	$\langle im_1, Simi_M^1 \rangle, \langle im_2, Simi_M^2 \rangle, \cdots, \langle im_{df_M}, Simi_M^{df_M} \rangle$

III. EXPERIMENT

We use WISAIAT to perform test both in a Corel dataset and a real Web image search dataset to demonstrate its effectiveness and efficiency.

A. *Retreival in the Corel Dataset*

We collect 5100 images from Corel photo gallery, which are classified into 40 categories using the terms like mountain, grass, horse, plane, bird, sky, firework, leaf, river, boat, etc.

Moreover, a thesaurus WordNet [16] is used to group synonymous query terms into a single concept. In the Corel dataset, 40x30 images (30 images for each category) are used for training, and then all the 5100 images are annotated. Retrieval test is performed both by 40 words and 510 images. The precision, recall and F-measure values when queried by the 40 words are listed in table II, while the average values of precision, recall and F-measure are 0.3110, 0.2810 and 0.2952 respectively when retrieved by the 510 images.

B. *Retrieval in the Yahoo! Dataset*

5260 images crawled from Yahoo! are used to evaluate the WISAIAT system. We classify all the 5260 images into 102 categories, choose 10 images for each category, and then train all the 1020 images. After that, all the 5260 images are automatically annotated by the rule output. Finally we implement the 102 word-search and random 500 image-search respectively. The average values or precision, recall and F-measure for Web image search are 0.4244, 0.1648 and 0.2374 respectively. Fig.4 gives an example of the search result by the query "pool". Fig.5 shows an image search result by selecting "precise image search" choice.

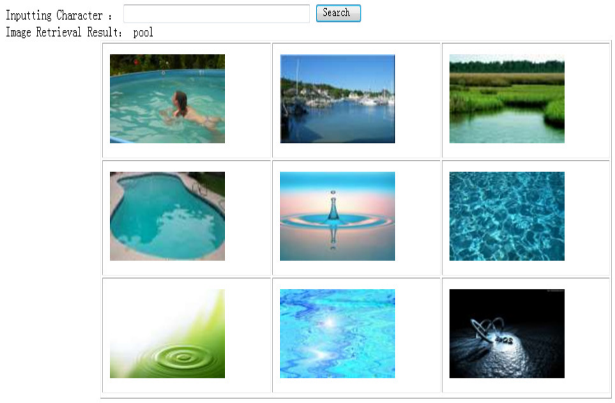

Image Retrieval Result: pool

Figure 4. Search result by the query "pool".

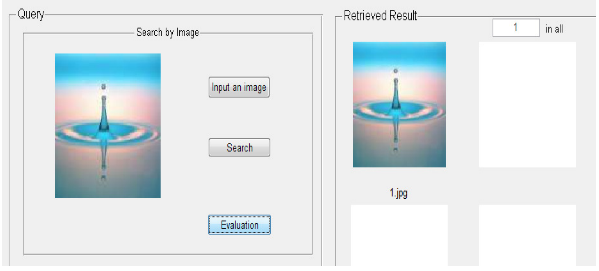

Figure 5. A precise image search result.

TABLE II. THE PRECISION, RECALL AND F-MEASURE VALUES FOR THE RETRIEVAL BY THE 40 WORDS

Terms	P	R	F	Terms	P	R	F
mountain	0.1157	0.556	0.1916	helicopter	0.0952	0.2	0.129
grass	0.0835	0.9	0.1528	building	0.1161	0.545	0.1914
horse	0.0297	0.94	0.0576	balloon	0.6487	0.24	0.3504
plane	0.0447	0.5	0.0821	light	0.1053	0.06	0.0764
bird	0.0797	0.395	0.1327	sunset	0.1808	0.81	0.2956
sky	0.1332	0.3143	0.1871	cloud	0.0362	0.5	0.0676
firework	0.0725	0.44	0.1245	tree	0.0987	0.77	0.175
night	0.0491	0.9	0.0931	sculpture	0.1628	0.14	0.1505
flower	0.0909	0.64	0.1592	house	0.1533	0.2625	0.1936
butterfly	0.0343	0.72	0.0655	monkey	0.0922	0.82	0.1658
cactus	0.0278	0.82	0.0537	sea	0.0433	0.5467	0.0803
car	0.3784	0.28	0.3218	boat	0.0851	0.16	0.1111
bonsai	0.1781	0.7	0.284	snow	0.0441	0.3	0.0769
elephant	0.5149	0.69	0.5897	road	0.0549	0.64	0.1011
bear	0.3623	0.5	0.4202	sand	0.0498	0.62	0.0922
wolf	0.2458	0.44	0.3154	rock	0.6905	0.58	0.6304
wave	0.0727	0.58	0.1292	stone	0.0959	0.14	0.1138
goat	0.2361	0.34	0.2787	leaf	0.2432	0.36	0.2903
tiger	0.6949	0.41	0.5157	river	0.534	0.3667	0.4348
bomber	0.8636	0.19	0.3115	ground	0.4058	0.0622	0.1079

IV. CONCLUSION

This paper presents a Web image search method by using automatic image annotation and translation. The prototype of the system can learn semantic image concepts from image content and translate unstructured images into textual documents, so that images are indexed and retrieved in the same way as textual documents, like text search. A new automatic annotation algorithm by combining DT with SVM is addressed. An inverted file is used to rank the search result instead of the conventional methods. Both word-search and image-search are performed and preliminary experiment results show the effectiveness and efficiency of the proposed methodologies. Continuous improvements to the system are being made. WISAIAT is expected to be more practical and serve as a real-time reranking system to the current Web image search result such as generated by Google or Yahoo!.

ACKNOWLEDGMENT

This research was supported in part by the scientific research foundation for the returned overseas Chinese scholars, state education ministry and in part by the specialized research fund for the doctoral program of higher education under Grant 20090184120022.

REFERENCES

[1] B. Smolka, M. Szczepanski, R. Lukac, and A.N. Venetsanopoulos, “Robust color image retrieval for the World Wide Web,” in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Poland, May 2004, vol. 3, pp. III-461-464.

[2] F. Jing, C.H. Wang, Y.H. Yao, K.F. Deng, L. Zhang, and W.Y. Ma, “Igroup: web image search results clustering,” in Proc. ACM Int. Conf. Multimedia, California, USA, Jun. 2006, pp. 377-384.

[3] R. Shi, C. Lee, and T. Chua, “Enhancing image annotation by integrating concept ontology and text-based bayesian learning model,” in Proc. ACM Int. Conf. Multimedia, Germany, Sep. 2007, pp. 341-344.

[4] Y. Liu, D.S. Zhang, and G.J. Lu, “Region-based image retrieval with high-level semantics using decision tree learning,” Pattern Recognition, vol. 41, pp. 2554-2570, Aug. 2008.

[5] J. Li and J.Z. Wang, “Automatic linguistic indexing of pictures by a statistical modelling approach,” IEEE Trans. PAMI, vol. 25, pp. 1075-1088, Sep. 2003.

[6] L.L. Cao and F.F. Li, “Spatially coherent latent topic model for concurrent object segmentation and classification,” in Proc. IEEE Int. Conf. Computer Vision, Rio de Janeiro, Brazil, Oct. 2007, pp. 1-8.

[7] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos, “Supervised learning of semantic classes for image annotation and retrieval,” IEEE Trans. PAMI, vol. 29, pp. 394-410, Mar. 2007.

[8] J. Li and J.Z. Wang, “Real-time computerized annotation of pictures,” IEEE Trans. PAMI, vol. 30, pp. 985-1002, Jun. 2008.

[9] J. Cui, F. Wen, and X. Tang, “Real time google and live image search re-ranking,” in Proc. ACM Int. Conf. Multimedia, Canada, 2008, pp. 729-732.

[10] Y. Jing and S. Baluja, “Pagerank for product image search,” in Proc. Int. Conf. World Wide Web, Beijing, China, Apr. 2008, pp. 307-316.

[11] H.C. Fu, Y.Y. Xu, and H.T. Pao, “Multimodal Search for Effective Image Retrieval,” in Proc. Int. Conf. Systems, Signals and Image Processing, Bratislava, Czechoslovakia, Jun. 2008, pp. 233-236.

[12] C. Wang, L. Zhang, and H. Zhang, “Learning to reduce the semantic gap in web image retrieval and annotation,” in Proc. Int. Conf. Research and Development in Info. Retrieval, Singapore, Jul. 2008, pp. 355-362.

[13] N. Hervé and N. Boujemaa, “Image annotation: which approach for realistic databases?,” in Proc. ACM Int. Conf. Image and Video Retrieval, Amsterdam, Netherlands, Jul. 2007, pp.170-177.

[14] Y. Lu, L. Zhang, Q. Tian, and W-Y. Ma, “What are the high-level concepts with small semantic gaps?,” in Proc. Int. Conf. Computer Vision and Pattern Recognition, Anchorage, USA, Jun. 2008, pp.1-8.

[15] D.S. Zhang, M. M. Islam, G. J. Lu, and J. Hou, “Semantic image retrieval using region based inverted file,” in Proc. Digital Image Computing: Techniques and Applications, Melbourne, Australia, Dec. 2009, pp.242-249.

[16] J. Yang, L. Wenyin, H. J. Zhang, and Y. Zhuang, “Thesaurus-aided approach for image browsing and retrieval,” in Proc. IEEE Int. Conf. Multimedia and Expo, Tokyo, Japan, Aug.2001, pp. 313-316.