

CBIR: From Low-Level Features to High-Level Semantics

Xiang Sean Zhou*, Thomas S. Huang

Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana Champaign, Urbana, IL 61801, USA
{xzhou2, huang}@ifp.uiuc.edu

ABSTRACT

The performance of a content-based image retrieval (CBIR) system is inherently constrained by the features adopted to represent the images in the database. Use of low-level features can not give satisfactory retrieval results in many cases; especially when the high-level concepts in the user's mind is not easily expressible in terms of the low-level features. Therefore whenever possible, textual annotations shall be added or extracted and/or processed to improve the retrieval performance. In this paper a hybrid image retrieval system is presented to provide the user with the flexibility of using both the high-level semantic concept/keywords as well as low-level feature content in the retrieval process. The emphasis is put on a statistical algorithm for semantic grouping in the concept space through relevance feedback in the image space. Under this framework, the system can also incrementally learn the user's search habit/preference in terms of semantic relations among concepts; and uses this information to improve the performance of subsequent retrieval tasks. This algorithm can eliminate the need for a stand-alone thesaurus, which may be too large in size and contain too much redundant information to be of practical use. Simulated experiments are designed to test the effectiveness of the algorithm. An intelligent dialog system, to which this algorithm can be a part of the knowledge acquisition module, is also described as a front end for the CBIR system.

Keywords: Content-based image retrieval; semantic grouping; automatic thesaurus generation; information retrieval.

1. INTRODUCTION

Multimedia database system is a very active research field in recent years, with the propelling forces being: 1. The exponential growth of multimedia information everyday as the data source; 2. The rapid expansion of the World Wide Web as the consumer market; 3. The seemingly unstoppable advances in the computer hardware industry as the technical support. Nevertheless, multimedia databases that can support fast and efficient indexing, browsing, and retrieval are still in short supply.

As an interdisciplinary research field, multimedia database research addresses and explores both database management issues and multi-dimensional signal/information processing techniques. There are many challenging research sub-areas, among which content-based image retrieval (CBIR) has been very active since early 90's³. However, despite the advances in both the feature selection techniques and matching and retrieval techniques, the current CBIR systems still have a major difficulty that it has yet to overcome, i.e., how do we relate the low-level features to the high level semantics? From our extensive experiments on CBIR systems using features like color, texture, shape, spatial layout, etc., and relevance feedback from the user, we still found out that the low-level contents cannot always describe the high level semantic concepts in the user's mind. This is one of the major burdens for a CBIR system to be implemented in practical image retrieval applications. To overcome this burden, on the one hand, one can go along the direction of searching for more low-level features that can improve the performance of the current content-based retrieval schemes; on the other hand, whenever possible, incorporation of text-based retrieval with content-based retrieval is desirable. Actually, most of the on-line commercial image databases are annotated with keywords or categories.

* Correspondence: Email: xzhou2@uiuc.edu; WWW: <http://www.ifp.uiuc.edu/~xzhou2>; Telephone: 217-244-2960; Fax: 217-244-8371

Keywords have direct mapping toward high-level semantics, but the mapping is not one-to-one: people use the same word for different meanings in different context, or use different words for similar or even the same concepts. Therefore a thesaurus is needed during the retrieval process, otherwise the performance of the keyword-based retrieval will be very limited, and depend upon the consistency of the annotation; and the consistency between the user and the annotation; and even the consistency among different users. One can use existing general-purpose thesauri in the system, but nevertheless a stand-alone thesaurus may contain too much redundant information that is not relevant to the database or the user's preference, which gives rise to the issue of *automatic thesaurus construction* from an image database.

In document processing literatures, there have been extensive research on how to automatically construct thesauri, but all of these are based on the statistical analysis on *term occurrences* and *co-occurrences* in a particular document collection^{1,14}. These techniques are not applicable in the case of image databases, where the annotation is usually in terms of a very small number of phrases or keywords, in which there is usually no *term co-occurrences*, i.e., semantically similar terms do not occur in the annotation of the same image. For example, the annotation of one image can contain "car", and for another image, "motorcycle"; but these two terms usually do not appear together in the annotation for one image, even though they are semantically similar in some sense.

In this paper we assume that some of the images in the database have textual annotations in terms of short phrases or keywords. These can come from pattern recognition¹⁰; automatic speech recognition, keywords spotting from text; or manual annotation, etc. We propose a statistical algorithm for semantic grouping of keywords based on user relevance feedback during the retrieval process. During each retrieval and user feedback process, the algorithm will dynamically update the weights in a semantic network consisting of the keywords appeared in the database. This algorithm is statistically effective and robust, and most importantly, it runs automatically in the background with little computational overhead. After extensive use of the database by the user, the output of the algorithm, i.e., the weights between pairs of terms will correspond to the "similarity" of the two terms, or the estimated probability for the user to request these two terms together in one query. By using Hopfield network or clique detection we can further group terms into semantic classes, which can be used to assist future retrieval process. Also since these "knowledge" are all extracted from the user feedback, the term association information can also be regarded as the user's "search habit" or "preference". Therefore, this real time thesaurus construction algorithm based on user feedback will provide a practical way for not only the semantic grouping of keywords but also learning of user preference.

The rest of the paper is organized as follows. Section 2 provides the related work in document processing domain and the system background for the algorithm proposed; In Section 3 we introduce the statistical algorithm for learning the semantic relations among keywords from user feedback; Section 4 compares two methods for semantic grouping based on the semantic network obtained in Section 3; Section 5 discusses the use of the obtained keyword classes to assist future retrievals. Conclusions are drawn in Section 6.

2. BACKGROUND

Research in document processing literatures provide various method of knowledge discover methods, most of them represent the knowledge with a semantic net, where the nodes of the network represent different types of concepts and the weighted links among the nodes indicates the relevance among the concepts^{14, 1, 2}. One form of weight computation is as follows²:

$$\text{Weight} (T_j, T_k) = \frac{\sum_{i=1}^n d_{ijk}}{\sum_{i=1}^n d_{ij}} \quad (1)$$

$$\text{Weight} (T_k, T_j) = \frac{\sum_{i=1}^n d_{ijk}}{\sum_{i=1}^n d_{ik}} \quad (2)$$

where d_{ij} (d_{ik}) is a boolean variable with values 0 or 1, indicates term T_j (T_k) in document i , and d_{ijk} indicates whether both term i and j are in document i . Most if not all of the weight computation techniques are based on the concepts of term frequency, document frequency, and inverse document frequency in a particular document collection. In image databases, it

is inadequate to directly adopt these co-occurrence based estimation methods, due to the lack of co-occurrence of semantically similar terms in the annotation of a single image. So we have to rely on a group of images, e.g., the set of feedback images from the user during the retrieval process, to estimate the relevance between keywords/terms for the automatic construction of thesauri. Nevertheless, The proposed algorithm can be regarded as a natural extension of the pseudoclassification techniques in the text-processing domain into the content-based image retrieval domain, with the differences not only in the application domains, but also in terms of how the relevance feedback is processed and whether it is a dynamic on-line process. For the proposed algorithm, the relevant and irrelevant images are considered jointly and in a computationally efficient way (instead of an iterative way¹⁴, which can be computationally expensive); and the weight adaptation is in real-time and dynamically follows the user retrieval preference. It is not just a once-for-all process as in the case of pseudoclassification techniques, whose effectiveness is usually questionable once outside the special cases in which (or outside the users for whom) they are generated¹⁴.

To use relevance feedback to facilitate the thesaurus generation, we need a system that can use low-level features to express high level semantics in a reasonable if not perfect way; that is, a subset of similar images in terms of semantics should appear together with certain probability, even just a small probability. Though this sounds to be a weak condition, it actually is not. The gap between low-level features and high-level semantics is yet to be filled and there is still a long way to go. To bridge the two, feature selection^{5,6,8,9,18} and on-line learning are two effective directions to explore. Our system uses edge/structural features¹⁸ in addition to color, texture, and layout features, and relevance feedback as the way of on-line learning from the user¹³. After the user submit a query, from the initial returns, the user can then select the images of interest to him/her by specify them as relevant. Based on this feedback the features are weighted accordingly to yield a better retrieval result in the next round. Our extensive experiments showed that with the extended feature vectors and the relevance feedback techniques, in most cases, a subset of semantically similar images would appear together in the return, thus made it possible for our proposed algorithm to perform the semantic weighting and grouping tasks.

3. SEMANTIC RELATIONS BETWEEN KEYWORDS

For an image database such as personal digital photo album, user can add text annotations either by hand or by an automatic speech recognizer. Or for a dynamic image database on the World Wide Web, keywords can be extracted from the surrounding or related text. Then keyword-based retrieval is possible. However, problems arise when different keywords, though semantically similar, are assigned to very similar images; or when the user failed to use the exact wording as the one used for the images in the database. Obviously, a thesaurus is needed to resolve term association problem. One option is to use stand-alone thesauri. But the major problem is that usually they are too large in size and can contain too much redundant information to be of practical use. Therefore we propose automatic thesaurus construction from user relevance feedback during the retrieval process. The assumption is that some images in the databases have keyword annotations. During the browsing or content-based retrieval process, the system can take relevance feedback from the user¹³, which is essentially the set of images user regards as “relevant” out of all the images shown to the user. Based on this information and the annotations for both the relevant and non-relevant images shown, a statistical formula can be applied to update the “similarity” or “closeness” among all the keywords assigned to the current images.

Among all the images shown, the *relevant set* is the set of images indicated by the user as “good” ones. If a term only appeared in the annotations for the images in the *relevant set*, it is called a *relevant term*. The number of occurrences of a *relevant term i* in the *relevant set* is called *relevant term frequency*, denoted as f_i . The number of co-occurrences of two relevant terms i and j in the same image is denoted by c_{ij} . The “relevance” of term i and j , S_{ij} , is then updated as

$$S_{ij} = S_{ij} + \max(f_i, f_j) \times (\min(f_i, f_j) - c_{ij}) \quad (3)$$

This formula is executed after every user feedback with more than one relevant images. Note that the above formula implies that if two terms appeared in the annotations for one images, we cannot get any information out of it. For example, a relevant image annotated as “car, house, tree...” provides us with very little information with regard to how related the concepts of “car” and “tree” are; but just that they happen to be in the same picture. One can argue that the term co-occurrence sometimes dose provide us with useful information because certain things are tend to be together, e.g., “beach” and “ocean”; while others are hardly together, e.g., “elephant” and “kitchen sink”. However we still believe that such co-occurrence information is far less important and less consistent in the image domain than in the video domain. In the video case, multiple cues can be used together to construct “multi-jects” and “multi-nets”¹⁰.

In a simulated experiment, 1000 images are randomly assigned 0-3 keywords from a set of 20 keywords. Then we perform n random browsing operations by the user. Assume whenever the user sees images of “cars” (row or column #3), or “truck”(#5), or “motorcycle”(#11), he/she will mark the image as “relevant”. Then the results of the updated “concept similarity matrix” is shown in figure 1, with the number of browsing operation $n = 5, 10, 30$, respectively. The images in figure 1 actually depict a matrix of weights between all pairs of concepts. The element in the matrix, m_{ij} , is the relevance measure between concept i and j , with a value normalized to within 0 and 1 for display. The bright dots indicate the higher weights between the concept #3 and #5, #3 and #11, #5 and #11. Since the relevance measure is symmetric, so the matrix is symmetric as well. Also the weight to the node itself (the diagonal elements) is set to be small. One can see that the similarity between the concepts #3, #5, and #11 is clearly captured even after a small number of operations.



Figure 1. Concept similarity matrix after 5, 10, 30 updates (first simulation) Figure 2. Concept similarity matrix (second simulation)

A more complicated simulation is carried out for 30 keywords. This time with probability 1/3, the user will mark “car”(#3), “truck”(#5), and “motorcycle”(#11) images as relevant whenever he/she sees them; and with probability 1/3, the user will mark “car”(#3) and “van”(#17) as they appear on the screen; and with probability 1/3, the user is searching for “tiger”(#15) and “lion”(#26) together. After 80 rounds of browsing and feedback, the concept similarity matrix is shown in figure 2. The bright dots clearly show the relevant concept pairs.

4. SEMANTIC GROUPING OF KEYWORDS

From the concept similarity matrix we can then either use a Hopfield network, or simply use a heuristic clique detection algorithm to obtain the semantic classes. The two methods are compared with the second simulation as the example.

To use the parallel activation scheme of the Hopfield network, the 30 concepts are treated as the nodes in the network and the weights m_{ij} shown in figure 2 are assigned as the synaptic weights between nodes². During the iteration, the output at node i at time $t+1$ is:

$$O_i(t+1) = \frac{1}{1 + \exp\left[\frac{s_i - in_i}{s}\right]} \tag{4}$$

where $in_i = \sum_{j=1}^{30} o_j m_{ij}$, $s_i (=0.3)$ is a bias, and $s(=0.1)$ serves to control the shape of the Sigmoid function. The convergence criterion is that the L_1 -distance between two adjacent output vectors less than a threshold, 0.001 in this paper.

If we activate the Hopfield net by assigning 1 to each of the 30 nodes and iterate four times for each activation, then the result is shown in table 1. If we iterate until converge, the results will be the same for all nodes: {3 5 11 15 17 26}. The reason is that we are using symmetric weighting, and the noisy estimation in the concept similarity matrix can spread any activation all over the network.

Table 1. Hopfield activation results with limited iterations

Concept #	Active nodes	Concept #	Active nodes	Concept #	Active nodes
1	3 5 11	11-“motorcycle”	3 5 11 17	21	3 5 11
2	3 5 11 17	12	3 5 11 17	22	3 5 11 17
3-“car”	3 5 11 17	13	3 5 11	23	3 5 11
4	3 5 11	14	3 5 11	24	3 5 11 17
5-“truck”	3 5 11 17	15-“tiger”	3 5 11 15 17 26	25	3 5 11 17
6	3 5 11	16	3 5 11	26-“lion”	3 5 11 15 17 26
7	3 5 11 17	17-“van”	3 5 11 17	27	3 5 11 17
8	3 5 11 17	18	3 5 11	28	3 5 11

9	3 5 11	19	3 5 11 17	29	3 5 11 17
10	3 5 11	20	3 5 11	30	3 5 11

It can be seen that concept 3, 5, 11 and 17 are one class and 15 and 26 are another. One may also notice that the Hopfield network is suitable for single-link classification system, i.e., within one class, each term is relevant to at least one other term in the same class. In this example, 17-“van” is only relevant to 3-“car”, but it is classified as a member of a bigger class. The drawback is obvious: if a user is only interested to “car” or “van”, as he/she has shown during the retrieval processes, the system has a hard time in isolating these two as a separate class.

On the other hand, a heuristic clique detection algorithm, though simple and runs the risk of hard-thresholding, can perform complete-link classification, i.e., each term is relevant to all other terms in the same class. In fact this is the very definition of a clique in graph theory. The results of this clique detection algorithm are three classes: {3, 5, 11}; {3, 17}; and {15, 26}. With complete-link classification, when the user query for “cars”, the system will have the “intelligence” to ask the user whether he/she is interested in {car, truck, motorcycle} or {car, van}.

5. AN INTELLIGENT RETRIEVAL SYSTEM

With the low-level features, the textual annotations, and the automatically generated thesaurus, a hybrid intelligent image retrieval system can be built to provide convenient retrieval for the user. A hybrid image object with metadata (both low-level features and keyword annotations) is depicted in Figure 1. The system is capable of conducting intelligent interaction with the user, e.g., on-line learning of the user preference/semantic grouping of keywords; intelligent dialog to understand the query and to guide the retrieval process; on-line feature selection (weighting) from relevance feedback. (See Figure 4.)

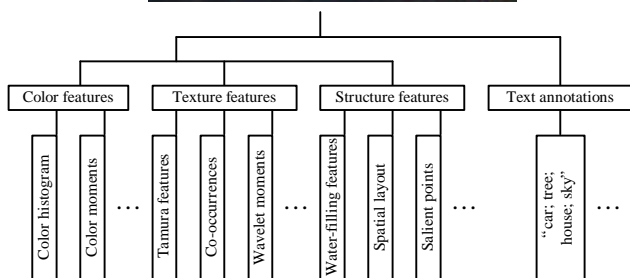


Figure 3. A hybrid image object

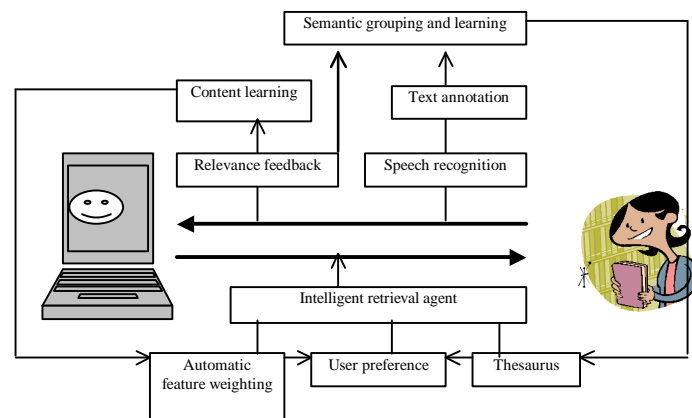


Figure 4. Human-machine intelligent interaction

6. CONCLUSIONS

A semantic grouping algorithm for keywords annotations based on the user feedback is presented and analyzed. This algorithm can eliminate the need for a stand-alone thesaurus. The system can incrementally learn the user’s search habit/preference in terms of semantic relations among concepts. The resulting semantic network and concept classes can be coupled with an intelligent agent based interface to assist the retrieval process. We propose this scheme in hope of bridging the gap between the machine and the human user for image retrieval system.

ACKNOWLEDGEMENTS

This work was supported in part by National Science Foundation Grant CDA 96-24386.

6. REFERENCES

1. Chen, H., B. Schatz, T. Ng, J. Martinez, A. Kirchhoff, C. Lin, "A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project", *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 18, No. 8, August, 1996
2. Chen, H., D. T. Ng, "An algorithmic approach to concept exploration is a large knowledge network", *J. Am. Soc. Information Science*, Vol. 46, No.5, pp.348-369, June 1995
3. Flickner, M. et al., "Query by image and video content: The qbic system", *IEEE Computers*. 1995
4. Gonzalez, R. C. and Woods, 1992, *Digital Image Processing*, Addison-Wesley
5. Haralick, R. M., K. Shanmugam, and I. Dinstein, "Texture feature for image classification", *IEEE Trans. SMC*, Nov. 1973
6. Hu, M. K., "Visual pattern recognition by moment invariants", *IRE Trans. Information Theory*, 8, 1962
7. Iqbal, Q. and J. K. Aggarwal, 1999, "Applying perceptual grouping to content-based image retrieval: Building images", *Proc. IEEE CVPR'99*, 42-48
8. Laine, A., J. Fan, 1993. "Texture classification by wavelet packet signatures". *IEEE Trans. Pattern Anal. Machine Intell.* 15, 1186-1191
9. Loupias, E. and N. Sebe, "Wavelet-based salient points for image retrieval", RR 99.11, *Laboratoire Reconnaissance de Formes et Vision*, INSA Lyon, Nov 1999. (<http://rfv.insa-lyon.fr/~louoias/points/>)
10. Naphade, M. R., T. Kristjansson, B. Frey, and T. S. Huang, "Probabilistic Multimedia Objects Multijets: A novel Approach to Indexing and Retrieval in Multimedia Systems" *Proc. IEEE ICIP*, Vol. 3 p. 536-540, Oct 1998, Chicago
11. Persoon, E. and K. S. Fu, 1977, "Shape discrimination using Fourier descriptors", *IEEE Trans. SMC*, Mar.
12. Ratan, A. L., et. al., 1999, "A framework for learning query concepts in image classification", *Proc. IEEE CVPR'99*, 423-429
13. Rui, Y., T. S. Huang, M. Ortega, and S. Mehrotra, 1998, "Relevance Feedback: A Power Tool in Interactive Content-Based Image Retrieval", *IEEE Tran on Circuits and Systems for Video Technology*, Vol 8, No. 5, Sept., 644-655
14. Salton, G., *Automatic Text Processing*. Reading, Mass.: Addison Wesley 1989
15. Smith, J. R. and Chang, 1995, "Transform features for texture classification and discrimination in large image databases", *Proc. IEEE ICIP'95*
16. Vailaya, A., A. K. Jain and H. J. Zhang, "On image classification: City Images vs. Landscapes", *Pattern Recognition*, vol. 31, December, 1921-1936, 1998
17. Zahn, C. T. and Roskies, "Fourier descriptors for plane closed curves", *IEEE Trans. Computers*, 1972
18. Zhou, X. S., Y. Rui, and T. S. Huang, "Water-filling: a novel way for image structural feature extraction", *Proc. ICIP*, 1999