AN UNSUPERVISED LEARNING APPROACH TO CONTENT-BASED IMAGE RETRIEVAL

Yixin Chen[1]*, James Z. Wang*[1,2]*, and Robert Krovetz*[3]

[1]Department of Computer Science and Engineering, [2]School of Information Sciences and Technology
[1,2]The Pennsylvania State University, University Park, PA 16802
[3]NEC Research Institute, 4 Independence Way, Princeton, NJ 08540

ABSTRACT

"Semantic gap" is an open challenging problem in content-based image retrieval. It reflects the discrepancy between low-level imagery features used by the retrieval algorithm and high-level concepts required by system users. This paper introduces a novel image retrieval scheme, CLUster-based rEtrieval of images by unsupervised learning (CLUE), to tackle the semantic gap problem. CLUE is built on a hypothesis that *images of the same semantics tend to be clustered*. It attempts to narrow the semantic gap by retrieving image clusters based on not only the feature similarity of images to the query, but also how images are similar to each other. CLUE has been tested using examples from a database of about $60,000$ general-purpose images. Empirical results demonstrate the effectiveness of CLUE.

1. INTRODUCTION

Developing an image searching and browsing algorithm, which can generate semantically accurate results, is an extremely difficult problem. However, with a single glance, human beings can tell the semantic similarity or difference between two images. This is probably because prior knowledge of similar images and objects may provide powerful assistance for humans in recognition. Can a computer program learn such knowledge or semantic concepts about images? In this paper, we attempt to address this question from the perspective of unsupervised learning.

1.1. Previous Work

In the past decade, many general-purpose image retrieval systems have been developed [14]. Examples include IBM QBIC System [5], MIT Photobook System [11], Berkeley Blobworld System [2], Virage System [6], Columbia VisualSEEK and WebSEEK Systems [15], the PicHunter System [4], UCSB NeTra System [9], UIUC MARS System [10], and Stanford WBIIS [18] and SIMPLIcity Systems [17].

A typical CBIR system views the query image and images in the database (target images) as a collection of features, and ranks the relevance between the query image and any target images in proportion to feature similarities. However, the meaning of an image is rarely self-evident. Images with high feature similarities to the query may be very different from the query in terms of the interpretation made by a user (*user semantics* or, in short, *semantics*). This is referred to as the *semantic gap*, which reflects the discrepancy between the relatively limited descriptive power of low level imagery features and the richness of user semantics.

Depending on the degree of user involvement in the retrieval process, two classes of approaches have been proposed to reduce the semantic gap: relevance feedback [4, 12] and image database preprocessing using statistical classification [1, 8, 16, 17]. Relevance feedback is effective for certain applications. Nonetheless such a system may add burden to a user especially when more information is required than just Boolean feedback (relevant or non-relevant). Statistical classification methods group images into semantically meaningful categories using low-level visual features so that semantically adaptive searching methods applicable to each category can be applied. Although these classification methods are successful in their specific domains of application, the simple ontology built upon them could not incorporate rich semantics of a sizable image database.

1.2. Motivation

Figure 1 shows a query image and the top 29 target images returned by a CBIR system described in [3] where the query image is on the upper-left corner. From left to right and top to bottom, the target images are ranked according to decreasing values of similarity measure. In essence, this can be viewed as one-dimensional visualization of image database in the "neighborhood" of the query image using a similarity measure. If the query image and majority of the images in the "vicinity" have the same user semantics, then we would expect good results. But target images with high feature similarities to the query image may be semantically quite different from the query image due to semantic gap. For the example in Figure 1, the target images belong to several semantic classes where the dominant ones include horses (11 out of 29), flowers (7 out of 29), golf player (4 out of 29), and vehicle (2 out of 29).

However the majority of top matches in Figure 1 belong to a quite small number of distinct semantic classes, which suggests a hypothesis that, in the "vicinity" of the query image, images tend to be semantically clustered in some feature space. Therefore, a retrieval method, which is capable of capturing this structural relationship, will be able to render semantically more meaningful results to the user than merely a list of images sorted by a similarity measure. This motivates us to tackle the semantic gap problem from the perspective of unsupervised learning. In this paper, we

The material was based upon work supported by the National Science Foundation under Grant No. IIS-0219272, the PNC Foundation, Penn State, NEC Institute, and SUN Microsystems

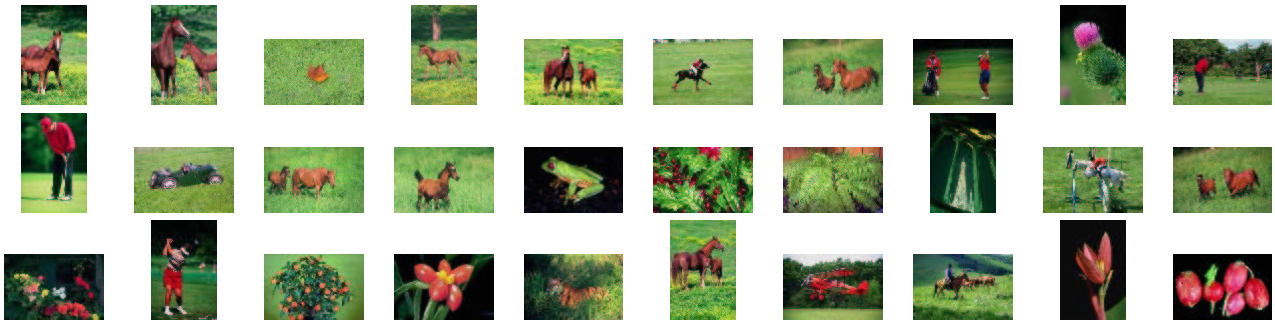

Figure 1: A query image and its top 29 matches returned by the CBIR system at `http://wang.ist.psu.edu/IMAGE` (UFM). The query image is on the upper-left corner. The ID number of the query image is 6275.

propose an algorithm, CLUster-based rEtrieval of images by unsupervised learning (CLUE), to retrieve image clusters instead of a set of ordered images: the query image and neighboring target images, which are selected according to a similarity measure, are clustered by an unsupervised learning method and returned to the user. CLUE has the following characteristics:

- It is a novel image retrieval scheme that attempts to reduce the semantic gap by providing image clusters, instead of a set of ordered images. The image clusters are obtained from an unsupervised learning process based on not only the feature similarity of images to the query, but also how images are similar to each other. In this sense, CLUE aims to capture the underlying concepts about how images of the same semantics are alike and present to the users semantic relevant *clues* as to where to navigate.
- It is a similarity-driven approach that can be virtually built upon any symmetric real-valued image similarity measure (metric or non-metric). Consequently, our approach could be combined with many other image retrieval schemes including the relevance feedback approach with dynamically updated models of similarity measure. Moreover it may also be used as a part of the interface for keyword-based image search engine.
- It provides a local visualization of the image database using a clustering technique. Because only images similar (close) to the query image in terms of a similarity measure are considered, the assumption of the simple semantically clustered structure may be reasonable. This is different to current image database statistical classification methods that try to represent the complex ontology of the whole image database using a simple structure.

2. RETRIEVAL OF IMAGE CLUSTERS

For the purpose of simplifying the explanations, we call a CBIR system using CLUE a Content-Based Image Clusters Retrieval (CBICR) system. From a data-flow viewpoint, a general CBICR system can be characterized by a diagram in Figure 2. The retrieval process starts with feature extraction. The features for target images (images in the database) are usually computed beforehand and stored as feature files. Using these features together with an

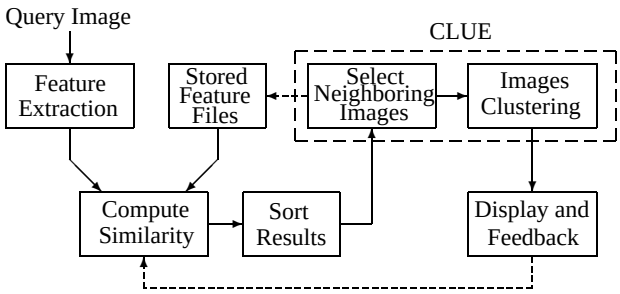

Figure 2: A diagram of a general CBICR system. The arrows with dotted lines may not exist for some CBICR systems.

image similarity measure, the resemblance between the query image and target images are evaluated and sorted. Next, a collection of target images that are "close" to the query image are selected as the neighborhood of the query image. A clustering algorithm is then applied to these target images. Finally, the system displays the image clusters and adjusts the model of similarity measure according to user feedbacks (if relevance feedback is included).

The major difference between CBICR and CBIR systems lies in the two processing stages, selecting neighboring target images and image clustering. A typical CBIR system bypasses these two stages and directly outputs the sorted results to the display and feedback stage. Figure 2 suggests that CLUE can be designed independent of the rest algorithmic components of the system because the only information needed by CLUE is the sorted similarities. This implies that CLUE may be embedded in a typical CBIR system regardless of the imagery features being used, the sorting method, and whether there is feedback or not.

To mathematically define the neighborhood of a point, we need to first choose a measure of distance. As to images, the distance can be defined by either a similarity measure (a larger value indicates a smaller distance) or a dissimilarity measure (a smaller value indicates a smaller distance). Because simple algebraic operations can convert a similarity measure into a dissimilarity measure, without loss of generality, we assume that the distance between two images is determined by a symmetric dissimilarity measure, $d(i, j) = d(j, i) \geq 0$, and name $d(i, j)$ the distance between images i and j to simplify the notation.

Neighboring images are selected by nearest neighbors method (NNM). It first chooses k nearest neighbors of the query image i as seeds. The r nearest neighbors for each seed are then found. Finally, the neighboring images are selected to be all the distinct images among seeds and their r nearest neighbors, i.e., distinct images in $k(r+1)$ target images.

Data representation is typically the first step to solve any clustering problem. In the field of computer vision, two types of representations are widely used. One is called the *geometric representation*, in which data items are mapped to some real normed vector space. The other is the *graph representation*. It emphasizes the pairwise relationship, but is usually short of geometric interpretation. When working with images, the geometric representation has a major limitation: it requires that the images be mapped to points in some real normed vector space. Overall, this is a very restrictive constraint. For example, in region-based algorithms [3, 7, 17], an image is often viewed as a collection of regions. The number of regions may vary among images. Although regions can be mapped to certain real normed vector space, it is in general impossible to do so for images unless the distance between images is metric, in which case embedding becomes feasible. Nevertheless, many distances for images are non-metric.

Therefore, this paper adopts a graph representation of images. A set of n images is represented by a weighted undirected graph $G = (\mathbf{V}, \mathbf{E})$: the nodes $\mathbf{V} = \{1, 2, \ldots, n\}$ represent images, the edges $\mathbf{E} = \{(i, j) : i, j \in \mathbf{V}\}$ are formed between every pair of nodes, and the non-negative weight w_{ij} of an edge (i, j), indicating the similarity between two nodes, is a function of the distance (or similarity) between nodes (images) i and j. Given a distance $d(i, j)$ between images i and j, we define $w_{ij} = e^{-\frac{d(i,j)^2}{s^2}}$ where s is a scaling parameter that needs to be tuned to get suitable locality property. The weights can be organized into a matrix $\mathbf{W}$, named the *affinity matrix*, with the ij-th entry given by w_{ij}. The same weight function has been used in [13].

Under a graph representation, clustering can be naturally formulated as a graph partitioning problem. Among many graph-theoretic algorithms, this paper uses the normalized cut (Ncut) algorithm [13] for image clustering. Compared with many other spectral graph partitioning methods, such as average cut and average association, the Ncut method is empirically shown to be relatively robust in image segmentation [13]. The Ncut method can be recursively applied to get more than two clusters. But this leads to the questions: 1) which subgraph should be divided? and 2) when should the process stop? In this paper, we use a simple heuristic. Each time the subgraph with the maximum number of nodes is partitioned (random selection for tie breaking). The process terminates when the bound on the number of clusters is reached or the Ncut value exceeds some threshold.

Ultimately, the system needs to present the image clusters to the user. Unlike a typical CBIR system, which displays certain numbers of top matched target images to the user, a CBICR system should be able to provide an intuitive visualization of the clustered structure in addition to all the retrieved target images. For this reason, we propose a two-level display scheme. At the first level, the system shows a collection of representative images of all the clusters (one for each cluster). At the second level, the system displays all target images within the cluster specified by a user. We define a representative image of a cluster to be the image that has the maximum sum of within cluster similarities.

3. EXPERIMENTS

Our experimental CBICR system uses the same feature extraction scheme and the UFM similarity measure as those in [3]. The system is implemented with a general-purpose image database (from COREL), which includes about $60,000$ images. The system has a very simple CGI-based query interface. It provides a *Random* option that will give a user a random set of images from the image database to start with. In addition, users can either enter the ID of an image as the query or submit any image on the Internet as a query by entering the URL of the image.

Qualitative performance evaluation of the system over COREL database is provided as follows. We randomly pick two query images with different semantics, namely, *birds* and *car* (more queries can be tested at our demonstration web site [1]. For each query example, we examine the precision of the query results depending on the relevance of the image semantics. Here only images in the first cluster, in which the query image resides, are considered. Since CLUE of our system is built upon the UFM similarity measure, query results of a typical CBIR system using the UFM similarity measure [3] (we call the system UFM to simplify notation) are also included for comparison (a demonstration for UFM is also available at our demonstration website). We admit that the relevance of image semantics depends on standpoint of a user. Therefore, our relevance criteria, specified in Figure 3, may be quite different from those used by a user of the system. Due to space limitations, only the top 11 matches to each query are shown in Figure 3. We also provide the number of relevant images in the first cluster (for CLUE) or among top 31 matches (for UFM).

Compared with the UFM, CLUE provides semantically more precise results for the query examples given in Figure 3. This is reasonable since CLUE utilizes more information about image similarities than the UFM does. CLUE groups images into clusters based on pairwise distances so that the within-cluster similarity is high and between-clusters similarity is low. The results seem to indicate that, to some extent, the unsupervised learning can group together semantically similar images.

4. CONCLUSIONS AND FUTURE WORK

This paper introduces CLUE, a novel image retrieval scheme, based on a simple assumption: semantically similar images tend to be clustered in some feature space. CLUE attempts to retrieve semantically coherent image clusters from unsupervised learning of how images of the same semantics are alike. It is a general approach in the sense that it can be combined with any real-valued symmetric image similarity measure (metric or non-metric). Thus it may be embedded in many current CBIR systems.

5. REFERENCES

[1] K. Barnard and D. Forsyth, "Learning the Semantics of Words and Pictures," *Proc. 8th Int. Conference on Computer Vision*, vol. 2, pp. 408–415, 2001.

[1] `http://wang.ist.psu.edu/IMAGE`(CLUE)

(a) 6 matches out of 11; 12 out of 29

3 matches out of 11; 9 out of 31

(b) 8 matches out of 11; 15 out of 26

4 matches out of 11; 7 out of 31

Figure 3: Comparison of CLUE and UFM. The query image is the upper-left corner image of each block of images. The underlined numbers below the images are the ID numbers of the images in the database. For the images in the left column, the other number is cluster ID. For images in the right column, the other two numbers are the value of the UFM measure between the query image and the matched image, and the number of regions in the image. (a) birds, (b) car. An image with a border around it is a representative image for a cluster.

[2] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image Segmentation Using Expectation-Maximization and its Application to Image Querying," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 8, pp. 1026–1038, 2002.

[3] Y. Chen and J.Z. Wang, "A Region-Based Fuzzy Feature Matching Approach to Content-Based Image Retrieval," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 9, pp. 1252–1267, 2002.

[4] I.J. Cox, M.L. Miller, T.P. Minka, T.V. Papathomas, and P.N. Yianilos, "The Bayesian Image Retrieval System, PicHunter: Theory, Implementation, and Psychophysical Experiments," *IEEE Trans. Image Processing*, vol. 9, no. 1, pp. 20–37, 2000.

[5] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, "Efficient and Effective Querying by Image Content," *J. Intell. Inform. Syst.*, vol. 3, no. 3-4, pp. 231–262, 1994.

[6] A. Gupta and R. Jain, "Visual Information Retrieval," *Commun. ACM*, vol. 40, no. 5, pp. 70–79, 1997.

[7] J. Li, J.Z. Wang, and G. Wiederhold, "IRM: Integrated Region Matching for Image Retrieval," *Proc. 8th ACM Int'l Conf. on Multimedia*, pp. 147–156, 2000.

[8] J. Li and J.Z. Wang, "Automatic Linguistic Indexing of Pictures By a Statistical Modeling Approach," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, 2003.

[9] W.Y. Ma and B. Manjunath, "NeTra: A Toolbox for Navigating Large Image Databases," *Proc. IEEE Int'l Conf. Image Processing*, pp. 568–571, 1997.

[10] S. Mehrotra, Y. Rui, M. Ortega-Binderberger, and T.S. Huang, "Supporting Content-Based Queries over Images in MARS," *Proc. IEEE Int'l Conf. on Multimedia Computing and Systems*, pp. 632–633, June 1997.

[11] A. Pentland, R.W. Picard, and S. Sclaroff, "Photobook: Content-Based Manipulation for Image Databases," *Int'l J. Comput. Vis.*, vol. 18, no. 3, pp. 233–254, 1996.

[12] Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra, "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval," *IEEE Trans. Circuits and Video Technology*, vol. 8, no. 5, pp. 644–655, 1998.

[13] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[14] A. W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[15] J.R. Smith and S.-F. Chang, "VisualSEEK: A Fully Automated Content-Based Query System," *Proc. 4th ACM Int'l Conf. on Multimedia*, pp. 87–98, 1996.

[16] A. Vailaya, M. A.T. Figueiredo, A.K. Jain, and H.-J. Zhang, "Image Classification for Content-Based Indexing," *IEEE Trans. Image Processing*, vol. 10, no. 1, pp. 117–130, 2001.

[17] J.Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 9, pp. 947–963, 2001.

[18] J.Z. Wang, G. Wiederhold, O. Firschein, and X.W. Sha, "Content-Based Image Indexing and Searching Using Daubechies' wavelets," *Int'l J. Digital Libraries*, vol. 1, no. 4, pp. 311–328, 1998.