MULTIDIMENSIONAL INTERACTIVE FINE-GRAINED IMAGE RETRIEVAL

Jieh Hsiang, Wen-Jun Liu, Bee-Chung Chen, and Hsieh-Chang Tu

National Taiwan University
Computer Science and Information Engineering
Taipei, Taiwan, R.O.C.

ABSTRACT

We propose an image retrieval methodology for a collection of similar images. By similar, we mean that one can define, for the collection, a set of *dimensions*, and for each of which a set of *features*. The dimensions are used to capture the essential characteristics of the images in the collection, and the features are for describing each image to a certain degree. We call this strategy *fine-grained image retrieval* to differentiate it from the more common *coarse-grained* retrieval, which does not assume any semantic properties on the image collection.

The effectiveness of our methodology is demonstrated through an icon-based interactive retrieval system on a collection of butterfly images. This system provides the user with a friendly initial *query-by-feature* (*QBF*) interface. The user can then use *query-by-example* (*QBE*) to refine the query.

In addition to presenting an outline of the methodology and the implementation on butterfly images, we also present some experimental results.

1. OVERVIEW

Image retrieval gained eminence in the age of the Internet due to the emergence of the vast variety of collections and wide range of users. A successful image retrieval system needs to capture the content information, recognize the differences between two images, and grasp the intention of the users. Most Internet image search engines are aimed at retrieving images available over the Web. Because they do not impose restrictions on the type of images, they cannot utilize structural and semantic information embedded in a collection of a specific domain. Although useful for Web retrieval, these *coarse-grained* retrieval systems cannot distinguish subtle differences between similar images.

In this paper we investigate a somewhat different problem. We propose a method for retrieving images from a specific domain. An example we will use to demonstrate our methodology is a collection of images of butterfly specimens. This type of problems is challenging because butterfly images are often different from each other in very subtle ways. A coarse-grained image retrieval system, not taking into consideration the information embedded in the collection, may return half of the butterfly data set for a given query. Another challenge is the impreciseness of human perception. For example, one person may describe a butterfly as a black one with white dots, while another may insist that the same butterfly is a white one with black stripes. A good retrieval system should be able to deal with such discrepancy. A third challenge is that it is not easy for a user to describe a butterfly to a retrieval system, especially at the beginning of a query session. (We call the starting query the *bootstrap* query.) Some image retrieval systems allow the user to input keywords as the bootstrap query, while others let the user inputs an image or draw some rough image outline. In our butterfly image retrieval system, we designed a very simple visual interface to let the user specify, pictorially, certain features of the intended butterfly.

The most fundamental assumption of our approach is a notion of *dimension*. That is, the images in the collection share certain similarity that can be captured by a set of dimensions. For instance, butterfly images can be characterized by their *color*, *shape*, and *pattern*. (Size is another possible dimension for butterflies, but it does not seem essential for our image retrieval need.) In another prototype system that we implemented for drugs, the dimensions are color, shape, pattern, and texture. For each dimension, a set of *features* is extracted from the images. The extraction can be done manually, semi-automatically, or automatically, depending on the complexity of the images. For the 1348 butterfly images in our collection, we extracted 18 color features, 7 shape features, and 17 pattern features.

We use a notion of *proximity* to measure the relationship between an image and a feature, and a notion of *similarity* to measure the relationship between two images. We designed some measures and compared their relative effectiveness.

The visual query interface works as follow: To initiate a query session, the system presents a page with an outline of a butterfly and menus of existing features of all the dimen-

The financial support of the National Science Council of R.O.C. under Grant NSC90-2213-E-002-141 is gratefully acknowledged.

sions (see Figure 1). The user starts the bootstrap query by selecting (at most) one feature from each dimension. The selected features will fill the butterfly outline, so that the user has a clear idea of what he has chosen. The bootstrap query is an important part of the *query-by-feature* (QBF) mechanism of our method. After the query is issued, the system returns a list of thumbnails of images that are closest to the features selected according to the proximity measure. The page that shows the query results is divided into two parts (see Figure 2). The left part is the list of thumbnails returned, and the results are summarized according to the features and listed, by the features, on the right. The user can issue another QBF by clicking on one of the features on the right. The user may also click on the link underneath any of the thumbnails to get a list of images that are closest to the one chosen (also arranged in two parts as in the QBF case). This retrieval mechanism is called *query-by-example* (QBE).

Our methodology has addressed several important issues in image retrieval. First we utilize the semantic information embedded, not only in the images, but also in the content domain itself. Second, our icon-based interface takes the impreciseness of the user perception into consideration. Most users have no idea how to describe a butterfly verbally. Our interface allows the user to provide a rough sketch of the image by selecting pre-defined features. The interactive feature of the query mechanism allows the user to explore further even if the bootstrap query is way off mark.

The notion of dimension is useful in all aspects of the image retrieval process. In the data processing phase, it captures the essential image characteristics of an entire collection, not just the images themselves; thus providing a simple way of classification for the image set. The use of dimensions also makes feature extraction and categorization easier. In the first implementation of our butterfly image retrieval system, we identified features (of each dimension) manually. Later we used this implementation as the *ground truth* to test the effectiveness of automatic clustering techniques. The use of dimensions breaks the feature extraction problem naturally into more manageable sub-problems, which makes it possible to find a reasonable size of features from clustering. Dimensions are also a powerful tool in the image retrieval phase. Our interface for the bootstrap query is essentially a multidimensional design, where features of the dimensions are shown as icons for the user to choose. The features of the dimensions are used in the post-classification of query results, which forms the basis of our interactive query mechanism.

Because our image retrieval methodology assumes that the target collection is from a specific domain and that we are aiming at the ability to separate subtle differences between images, we term this type of retrieval *fine-grained* image retrieval, to differentiate it from the *coarse-grained* methods that work on general collections.

The rest of the paper is organized as follows. In Section 2 we give some background information on image retrieval. We then present, in Section 3, an outline of our methodology. We also describe the butterfly image retrieval system that we mentioned earlier. Some experimental results are shown in Section 4, followed by concluding remarks and future work in Section 5.

2. BACKGROUND

Image retrieval can be generally divided into two layers: the *structural* and *semantic* layers. In the structural layer, the information of a raw image is represented as a descriptor of distinct *primitive-features* (such as color moments and shape fourier descriptor.) The *similarity* relationship between two raw images is measured by matching their descriptors. In the semantic layer, the information of an image is analyzed and then represented as a descriptor which may be a set of keywords or a semantic structure such as tree or graph. The similarity between two images is measured by matching their descriptors.

Regardless of the layer in which the retrieval system is designed, the ultimate goal is to retrieve images that the user wants. A user's search need is often unclear at the beginning of a search process, thus a good system should be able to guide the user and finally satisfy his search needs. The interactive nature of image retrieval, from a psychological point of view, was investigated in Perry et al [1].

Working at the structural layer has the clear advantage of having the ability to compute the descriptor automatically. The query results from such an approach, however, usually do not conform to the perception of the user. Smeulders et al [2] suggested that researchers in computer vision should focus more on identifying features required for interactive image understanding than on automatic techniques. Catarci [3] also discussed examples to show the importance of human-computer interaction. All these studies pointed out the importance of an effective and interactive visual interface to the success of an image retrieval system.

The differences in image retrieval for generic domain and specific domains were discussed in [2]. How to utilize the semantic properties embedded in a specific domain to aid the data construction, retrieval, and user interaction aspects of an image retrieval process is the goal of our paper.

3. A METHODOLOGY FOR FINE-GRAINED IMAGE RETRIEVAL

Our methodology includes two parts: the *data model* describes the essential data, while the *query model* includes

the descriptions of QBF, QBE, and interactive query facilities.

3.1. Data Model

- A collection of 1348 images of butterfly specimens: A
- A set of dimensions: $D = \{c, s, p\}$, (i.e. "color", "shape", "pattern")
- A set of features: $F = F_c \cup F_s \cup F_p$, where F_d is a set of features of dimension d.
- The proximity data on $A \times F$ and similarity data on $A \times A$ are measured by the following proximity and similarity functions, respectively.

$$\mathcal{P}(a, f) \geq 0, \text{ where } a \in A \text{ and } f \in F$$
$$\mathcal{S}(a, b) \geq 0, \text{ where } a, b \in A$$

3.2. Query Model

Figure 1 illustrates an example of a QBF query. This figure indicates that the user selected a feature "two horizontal bands" of dimension "pattern" and "red orange" of "color". For such a query q, the query results, obtained from the query function $\mathcal{QBF}(q, A)$ according to the proximity measure, are displayed on the left of Fig. 2 and categorized by distinctive features of "shape" on the right.

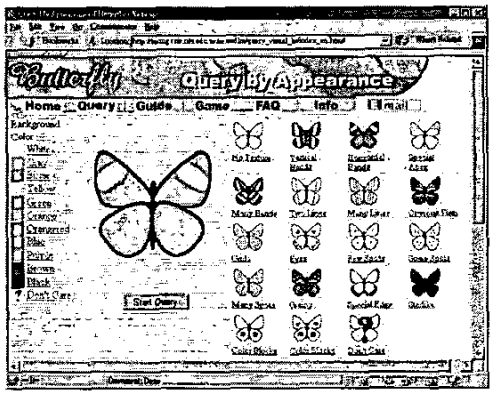

Fig. 1. Bootstrap interface and a QBF query example.

In Fig. 2, the user can then issue a QBE query to refine the query by clicking on one of the "Similar to this" links. The query function $\mathcal{QBE}(q, A)$ takes as arguments a query q and a set of images A, and then returns the images similar to the chosen one specified in q. Figure 3 illustrates the conceptual diagram of human-computer interaction.

4. EXPERIMENTS AND RESULTS

The manually-built butterfly image retrieval system is effective but labor-expensive. Our goal here is to construct

Fig. 2. QBF query results and QBE.

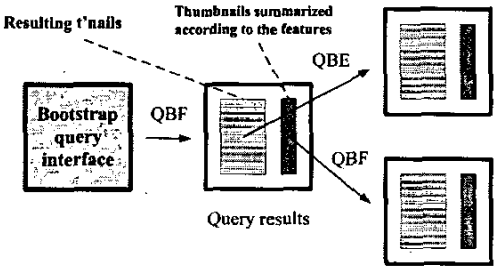

Fig. 3. Conceptual diagram of QBF/QBE.

similarity automatically and to make the constructed data conformable to human perception. We designed some measures and used the system as the ground truth to compare their relative effectiveness.

4.1. Experimental Methods and Data

We represented each of the images as a vector of primitive features or of our predefined features, and used different distance functions to compute $\mathcal{S}(a, b)$. Adopted primitive-features include 12 moments [4] in HSV color space for color, 13 CSS coefficients [5] for shape, and 60 energies of Gabor Wavelet [5] for pattern. We called this data set *primitive-features with real values* (**PFR**):

$$\mathbf{PFR} = \{\mathbf{v}_a \in \mathbb{R}^{85} : a \in A\}$$

In addition to **PFR**, we manually tagged $\mathcal{P}(a, f)$ as one of the four values: "very match", "match", "somewhat match", and "not match". These four levels are simply quantized as 3, 2, 1, and 0, respectively. We called this data set *features with level-values* (**FL**).

$$\mathbf{FL} = \{\mathbf{v}_a = (v_1, \cdots, v_{42}) : v_i \in \{3, 2, 1, 0\}\}$$

The vector size is 42 since there are 18 color features, 7 shape features, and 17 pattern features for the butterfly re-

trieval system. **FL** is also used to measure $S(a,b)$. This aims to evaluate the effectiveness of predefined features and human assistance.

Two distance functions, the Euclidean distance (**ED**) and an optimization scheme (**RF**) [6], are used for $S(a,b)$. The latter is a two-steps procedure as follows. Given an artificial image q and an image a, the similarity values of a to each of the images are computed from: (1) use the Lagrange multiplier to minimize the distance between q and a and to update the weights used in **RF**, and (2) use the new weights to compute the distances between a and each image.

We experimented with utilizing different amount of components of vectors of **PFR** (or **FL**) to test if more components (information) will result in higher precision for similarity.

4.2. Evaluation

For an image a, the ground-truth similarity data of a to each image are ranked and divided into four levels, i.e. "very similar", "similar", "somewhat similar", and "not similar". The computed similarity data are also ranked. Let Δ be the set of images in the first two levels of a ground-truth sequence and Ψ the set of the first $|\Delta|$ images of the corresponding experimental sequence, thus the *R-Precision* is computed as

$$\text{R-Precision} = |\Psi \cap \Delta| / |\Delta|.$$

4.3. Experimental Results

In Fig. 4, the "C-S-P", "C-P-S", and "C1-S1-P1" in all plots are different ways to increase the vector size. "C-S-P" denotes the order of "color", "shape", and "pattern"; and "C1-S1-P1" indicates increasing by one respectively from three dimensions.

In each plot of Fig. 4, the x-axis represents "dimensionality", while the y-axis "average R-Precision". It is obvious that the average R-Precision increases steadily in the plot (a) and (b) as increasing the vector size but behaves unsteadily in (c) and (d). The effectiveness of **FL** shown in this figure suggests a way to compute similarity automatically by **FL** with Euclidean distance. When registering a set of new images N, the data set **FL** of N can be computed from **PFR**, followed by manual modification. The modified proximity data are then used to compute similarity data automatically.

5. CONCLUSION AND FUTURE WORK

We propose a dimension-based methodology for fine-grained image retrieval. The use of dimensions can break the feature extraction problem into more manageable sub-problems, and the use of features can facilitate the construction of similarity from the experimental results. The notion of dimension can also be used to design an intuitive bootstrap interface.

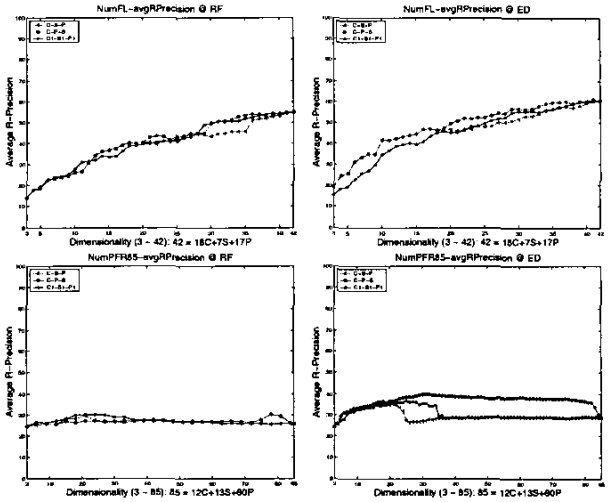

Fig. 4. Plot (a) **FL+RF** (upper-left); (b) **FL+ED** (upper-right); (c) **PFR+RF** (lower-left); and (d) **PFR+ED**.

The features used in the post-classification of query results also facilitates the human-computer interaction.

Two issues have not been discussed in this paper: how to add a new image and how to add a new feature that may result from adding a new image. Methods for solving these problems will be discussed in the full version of this paper.

6. REFERENCES

[1] B. Perry, S.-K. Chang, J. Dinsmore, D. Doermann, A. Rosenfeld, and S. Stevens, *Content-based Access to Multimedia Information: From technology trends to state of the art*, Kluwer Academic publishers, Norwell, Massachuset, 1999.

[2] 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. On Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, Dec. 2000.

[3] T. Catarci, "Strategic directions in human computer interaction," *ACM Computing Surveys*, vol. 28, no. 4, 1996.

[4] J. P. Eakins, *Lectures on Information Retrieval*, chapter Retrieval of Still Images by Content, pp. 110–138, ESSIR 2000. Springer-Verlag, Heidelberg, Berlin, Germany, Sept. 2000.

[5] MPEG7 committee, "MPEG7: Multimedia content description interface - part 3 visual," ISO/IEC JTC1/SC29/WG11/N4062, MPEG, March 2001.

[6] Y. Rui and T. S. Huang, "A novel relevance feedback technique in image retrieval," in *Procs. of the 7th ACM International Conference (part 2) on Multimedia*, 1999.