A Content-Based Image Retrieval System for Fish Taxonomy

Yixin Chen*
Dept. of Computer Science
University of New Orleans
New Orleans, LA 70148, USA
yixin@cs.uno.edu

Henry L. Bart, Jr.†
Tulane University Museum of Natural History
Belle Chasse, LA 70037, USA
hank@museum.tulane.edu

Fei Teng
Dept. of Computer Science
University of New Orleans
New Orleans, LA 70148, USA
fteng@cs.uno.edu

ABSTRACT

It is estimated that less than ten percent of the world's species have been discovered and described. The main reason for the slow pace of new species description is that the science of taxonomy, as traditionally practiced, can be very laborious: taxonomists have to manually gather and analyze data from large numbers of specimens, often from broad geographic areas, and identify the smallest subset of external body characters that uniquely diagnoses the new species as distinct from all its known relatives. The pace of data gathering and analysis in taxonomy can be greatly increased by the development of information technology. The Internet is being used to link taxonomists, taxonomic literature and specimen databases in different parts of the globe, and hence enables the development of tools for remote study of specimens archived as digital images. In this paper, we propose a content-based image retrieval system for taxonomic research. The system has a learning component that can identify representative body shape characters of known species based on digitized landmarks. The system can also provide statistical clues for assisting taxonomists to identify new species or subspecies. The experiments on a taxonomic problem involving species of suckers in the *Carpiodes* genus demonstrate promising results.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.5 [**Pattern Recognition**]: Applications

*Dr. Chen is also with the Research Institute for Children, 200 Henry Clay Avenue, New Orleans, LA 70118 USA.

†Dr. Bart is also with the Department of Ecology and Evolutionary Biology, Tulane University, New Orleans, LA 70118, USA.

MIR '05, November 10-11, 2005, Singapore.

Keywords

Content-based image retrieval, shape analysis, feature selection, image classification, taxonomic research

1. INTRODUCTION

Approximately 1.4 million species are known to science. However, estimates based on the rate of new species discovery place the total number of species on earth about 10-30 times of this number. Most unrecognized species are in poorly studied groups (e.g., insects) occurring in unexplored habitats (e.g., remote tropical forests). However, a surprising number of new species are still being discovered in developed countries with long histories of taxonomic research. Human population expansion and habitat destruction are causing extinctions of both known and yet to be discovered species. The accelerated pace of species decline has fueled the current *biodiversity crisis* [20], in which it is feared that many of the earth's species will be lost before they can be discovered and described.

1.1 The Science of Taxonomy

The job of discovering and describing new species falls on specialists in an area of biology called *taxonomy*. The science of taxonomy has been suffering from dwindling numbers of experts over the past few decades [21]. Moreover, the pace of taxonomic research, as traditionally practiced, is very slow. To describe new species of animals, taxonomists follow precise rules laid out in the International Code of Zoological Nomenclature [13]. They first have to recognize the species as distinctive from other know species (alpha taxonomy). They then demonstrate the distinctiveness of the species by comparing it to closely related species (beta taxonomy). Finally, they apply a name to the new species (gamma taxonomy). In recognizing the species as new, taxonomists use a *gestalt* recognition system that integrates multiple characters of body shape, external body characteristics, and pigmentation patterns. They make careful counts and measurements on large numbers of specimens from multiple populations across the geographic ranges of both the new and closely related species. The goal of data analysis is to identify the smallest subset of external body characters that uniquely diagnoses the new species as distinct from all of its known relatives. The process is laborious and can take years or even decades to complete, depending on the geographic range of the species.

The pace of data gathering and analysis in taxonomy can

be greatly increased by the development of information technology. The establishment of the Internet has brought forth a revolution in information storage, distribution, and processing. The World-Wide Web is being used to link taxonomists, taxonomic literature and specimen databases in different parts of the globe, and hence enables the development of tools for remote study of specimens archived as digital images [29].

A family of software tools has been designed in recent years for gathering and analyzing data on shape variation from images of specimens [22, 4, 31]. These software tools, referred to collectively as *geometric morphometrics* software, use homologous landmarks (points that are arguably related by evolutionary descent) along the body and referenced to (x, y) coordinates. The software allows superposition and alignment of landmarks from different specimens, adjustment for body size differences among specimens, and multivariate statistical analysis of derived shape variables. These analyses can only help taxonomists recognize overall shape differences among specimens.

However, none of the current software tools supports efficient searching and navigating through large image databases of specimens. Development of an effective image retrieval system would provide taxonomists with a powerful research tool and would allow them to pursue taxonomic research in a distributed environment. For example, a fish taxonomist in California wants to determine whether a recently captured specimen belongs to a new species by comparing it with specimens archived in a natural history museum at a distant location, e.g., the Tulane University Museum of Natural History (TUMNH) in Belle Chasse, Louisiana. The TUMNH Fish Collection is the largest collections of post larval fishes in the world[1]. It presently contains well over 7 million specimens, some of which have been digitally archived. If the TUMNH database had an image search engine linked to the Internet, the fish taxonomist in California could easily access the online system and retrieve images of specimens of interest and the relevant collection event information needed for the determination.

1.2 Related Work in Image Retrieval

Image retrieval algorithms roughly belong to two categories, depending on the query format: text-based approaches and content-based methods. The text-based approaches are based on the idea of storing a keyword, a set of keywords, or a textual description of the image content, created and entered by a human annotator, in addition to a pointer to the location of the raw image data. Image retrieval is then shifted to standard database management capabilities combined with information retrieval techniques.

Content-based image retrieval (CBIR) methods search images based on information automatically extracted from pixels. Initially, researchers focused on querying by image example, where a query image or sketch is given as input by a user [10, 19, 25, 8, 12, 17, 18, 7, 11, 28, 5, 6, 16]. Later systems incorporated feedback from users in an iterative refinement process [27, 32, 9]. From a computational perspective, a typical CBIR system views the query image and images in the database (i.e., target images) as a collection of features. It ranks the relevance between the query and any target image in proportion to a similarity measure calculated from the features. In this sense, these features, or signatures of images, characterize the content of images. And the similarity measure quantifies the resemblance in content features between a pair of images [23]. Readers are referred to [24] for a more comprehensive review of CBIR.

[1] `www.museum.tulane.edu`

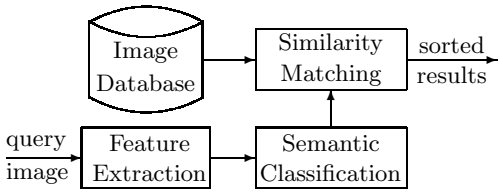


Figure 1: The structure of the CBIR system.

1.3 An Overview of Our System

The nature of taxonomic research brings the following requirements to the design of an image retrieval system:

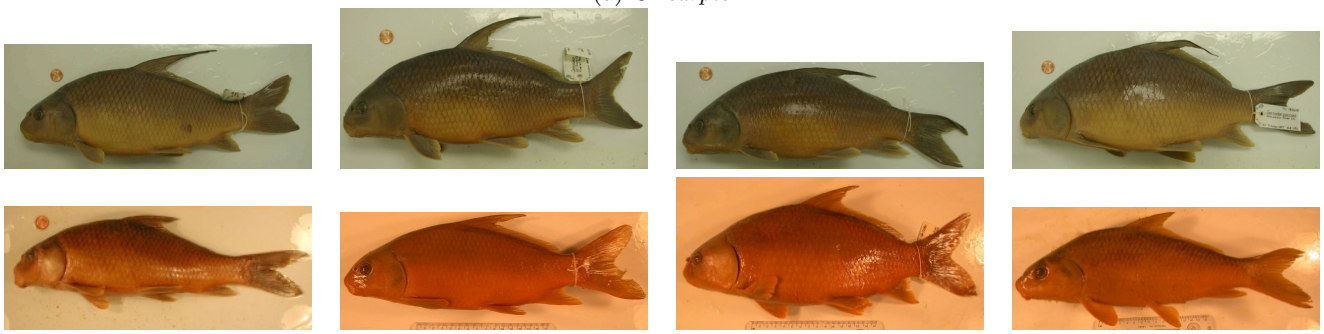
- *Text query*: Images of specimens from a natural history museum (i.e., the image database) almost always have textual annotations, e.g., location and date of capture, size of specimen, species, etc. Therefore, the image retrieval system should support text-based searches.
- *Query by example*: A typical usage scenario of the system is to find specimens in the database that are "semantically similar" to the query specimen based on the query image. This is clearly a query by example situation. From a taxonomic research point of view, the image *semantics* is defined as groupings of related specimens at different hierarchical levels, which, in the science of taxonomy, are referred to as taxa of varying rank, i.e., families, genera, species complexes, species, and subspecies.
- *Learning component*: For the query by example process, the system needs certain mechanisms to associate feature similarity with semantic similarity, i.e., bridging the *semantic gap*. One possible way is to include a learning component capable of identifying the feature characters that unite populations within each semantic class as well as distinguishing among semantic classes.

In this paper, we focus on the CBIR part of the system. Specifically, we propose a computational framework for categorizing semantic classes of populations based on body shape features, and retrieving images of specimens accordingly. The proposed framework can benefit the taxonomic research in the following ways:

- It provides taxonomists a tool of efficient searching, browsing, and retrieving images of specimens archived in natural history museums at distant locations.
- It automatically identifies an "optimal" set of body characters that unites populations within species, as well as distinguishes among species. Hence it can provide statistical clues in assisting the discovery of new species or subspecies.

As shown in Figure 1, the system has three major components: feature extraction, semantic classification, and similarity matching. In Section 2, we introduce background information on a taxonomic problem in the fish genus *Carpiodes*. Section 3 describes the feature extraction process.

(a) *C. carpio*



(b) *C. cyprinus*

(c) *C. velifer*

Figure 2: Images of specimens from three species of the genus *Carpiodes*: *C. Carpio*, *C. cyprinus*, and *C. velifer*.

Section 4 presents a joint feature selection and classification approach for semantic classification based on 1-norm support vector machines (SVMs). Section 5 describes a similarity matching scheme based on the distance in the overall shape space and semantic classification. Section 6 demonstrates extensive experiments on a data set of *Carpiodes* and discusses the results. Conclusions and possible future work are given in Section 7.

2. BACKGROUND KNOWLEDGE

The image database used in this paper comprises digital photographs of suckers of genus *Carpiodes*. However, our approach can be applied to any fish populations. The genus *Carpiodes*, as currently recognized, comprises three widely distributed species: the river carp-sucker *Carpiodes carpio* (*C. carpio*); the quillback *Carpiodes cyprinus* (*C. cyprinus*), and the highfin carp-sucker *Carpiodes velifer* (*C. velifer*). Figure 2 shows representative images of specimens of the three species. Most taxonomists regard each of these species as a complex of multiple biological species in need of revision [26]. The goal of the *taxonomic revision* in this case is to identify and formally describe the unrecognized species.

Over the past decade, geometric morphometric techniques have been developed for analyzing variation in body shape using a collection of coordinates of biologically definable, homologous landmarks along the body outline [1]. Figure 3 shows 15 homologous landmarks digitized on a specimen using the TpsDIG software tool developed by F. James Rohlf of SUNY Stony Brook[2]. The analysis methods accompanying the software focus on the landmark coordinates and geometric information about their relative positions. Through the alignment of landmarks and statistical analysis of the derived shape variables, groups of specimens may be identified as distinct in overall shape space. Unfortunately, the current geometric morphometric methods have two major limitations that hinder successful applications in taxonomic revision tasks:

[2]`http://life.bio.sunysb.edu/morph/`

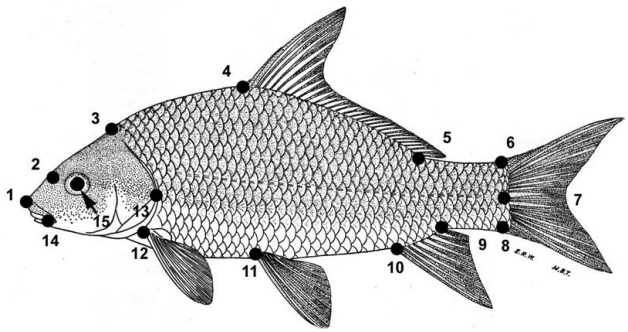


Figure 3: Digitized 15 homologous landmarks using TpsDIG Version 1.4 (2004 by F. James Rohlf).

- Groups of specimens are distinguished from other populations based on a small set of derived variables, which are usually functions (in their simplest form, linear combinations) of all shape variables. As such, derived variables are difficult to interpret in terms of particular body characters that taxonomists commonly use in diagnosing new species.

- Shape variation of specimens from closely related species or subspecies may not be discernible in overall shape space or through analysis of landmark coordinates. Therefore, current geometric morphometric methods may generate misleading results (see the example to be presented next).

Over the years since [26] was published, one of us (HLB) has examined shape and DNA sequence variation in all *Carpiodes* populations. Figure 4 shows the results of an analysis of overall body shape based on a geometric morphometric technique using canonical variate analysis (CVA). CVA grouped specimens from the Rio Grande (squares), upper Colorado River (stars), and other western Gulf Slope rivers with *C. carpio* specimens (circles) from the Mississippi River Basin. However, a surprising finding from the DNA sequence analysis was that the forms in Rio Grande and upper Colorado River system of Texas do not agree at all with *C. carpio*. Rather, they are closely related to *C. cyprinus*, which was not known to occur on the western Gulf Slope. Careful inspection of *Carpiodes* specimens in the Rio Grande and upper Colorado River system reveals that they lack the protuberance ("nipple") on the lower lip, which is diagnostic of *C. carpio* and *C. velifer*. They also have a relatively large head and a long snout, characters seen only in *C. cyprinus*. However, specimens from these populations also have an elongate and slender body, and it is these characters that cause them to be erroneously classified as *C. carpio* based on overall body shape analysis.

It took HLB three years of careful study of over 1000 *Carpiodes* specimens to determine that Rio Grande and upper Colorado River populations were misdiagnosed as *C. carpio*, and instead represented a new species related to *C. cyprinus*. The question we address next is: Can CBIR based on shape features be applied in a way that diagnoses taxonomic groups in genus *Carpiodes* more quickly and accurately?

3. FEATURE EXTRACTION

We focus on the digitized images of specimens with landmarks specified as in Figure 3. Let $LM_j \in \mathbb{R}^2$, $j = 1, \cdots, 15$,

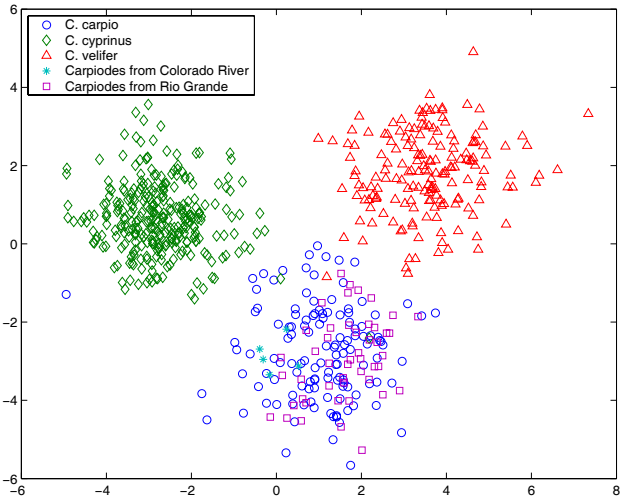


Figure 4: Plot of 650 *Carpiodes* specimens representing three distinct morphotypes on the first two canonical variate axes based on derived shape variables from geometric morphometric analysis of landmark data.

be the coordinates of landmarks on a specimen. We used the technique of Generalized Procrustes Analysis [14] to remove non-shape related variation in landmark coordinates. Specifically, the centroid of each configuration (based on the 15 landmarks associated with each specimen) was translated to the origin, and configurations was scaled to a common unit size.

We computed 12 features, $x_1, \cdots, x_{12}$, for each specimen using the 15 landmarks. These features correspond to different shape characters that taxonomists use to describe species. The description of each feature is given in Table 1. These features are divided into two groups:

- x_1–x_7: They describe shape characters that can be easily identified visually, for example, the size of head, the length of body, the distance between the tip of the snout and the nostril, the size of head in proportion of body size, etc.

- x_8–x_{12}: They can be easily evaluated from the landmark coordinates, but may not have a straightforward visual interpretation. These are the features that a domain expert may not identify easily, but are candidates of good indicators.

All 12 features were normalized across the specimens via translation and scaling.

4. SEMANTIC CLASSIFICATION

Semantic classification in our CBIR system targets the following taxonomic problem: given a collection of labeled specimens ($\mathbf{x}_i$'s) represented in a feature space, identify features and construct classifiers based on the selected features to distinguish among the known categories (or species). This problem is closely related to taxonomic revision: if the classifiers indeed capture the shape properties describing the known species, the classifiers will be helpful in discovering new species whenever there is shape variation between the new species and all the known species. Next, we describe a scenario for new species detection. Given a collection of

Table 1: Features describing shape characters. Non-shape related variation has been removed from LM_i, the landmark coordinates.

x_1	The distance between the tip of the snout and the naris, computed as the distance between LM_1 and LM_2.
x_2	The slope of the line connecting the tip of the snout and the naris, computed as the angle between the vertical axis and the line connecting LM_1 and LM_2.
x_3	The distance between the naris and the back of the mouth, computed as the distance between LM_2 and LM_{14}.
x_4	The slope of the line connecting the naris and the back of the mouth, computed as the angle between the vertical axis and the line connecting LM_2 and LM_{14}.
x_5	The size of head in proportion of the size of the body, computed as the area of the head polygon (vertices defined in sequence by LM_1, LM_2, LM_3, LM_{13}, LM_{12}, and LM_{14}) divided by the area of the body polygon (vertices defined in sequence by LM_3, LM_4, LM_5, LM_6, LM_7, LM_8, LM_9, LM_{10}, LM_{11}, LM_{12}, and LM_{13})
x_6	The length of the head in proportion of the length of the body, computed as the distance between LM_1 and LM_{13} divided by the distance between LM_{13} and LM_7.
x_7	The distance between LM_7 and LM_8.
x_8	The sum of the distance between LM_3 and LM_{13}, the distance between LM_{12} and LM_{13}, and the distance between LM_1 and LM_{13} divided by the distance between LM_{13} and LM_7.
x_9	The distance between the naris and the tip of the snout in proportion to the distance between the naris and the eye, computed as the distance between LM_1 and LM_2 divided by the distance between LM_2 and LM_{15}
x_{10}	The distance between LM_4 and LM_{11} divided by the distance between LM_{13} and LM_7.
x_{11}	The distance between LM_3 and LM_4 divided by the distance between LM_{13} and LM_7.
x_{12}	The angle between the vertical axis and the line connecting LM_{10} and LM_5.

specimens (not just a single specimen) from the same, but unknown, population, if the classifier assigns the specimens to several species without any preference on any one of the species, it is likely that the specimens are from a new species in need of description. If the classifier assigns the majority of the specimens to a single species, it is likely that the specimens belong to that known species. The underlying assumption is that specimens of a new species would *confuse* the classifier built upon all known species.

The classification of $\mathbf{x}_i$ is clearly a multi-class problem. We propose to use a tree structure to organize binary classifiers into a multi-class classifier. For example, Figure 5 shows a hierarchical classifier consisting of two binary classifiers for the identification of all three known species in *Carpiodes* genus. Finding an "optimal" structure is an interesting research topic for its own sake, but is beyond the scope of this paper. Here we assume the structure is determined beforehand. It is worth mentioning the difference between the classification tree in Figure 5 and the taxonomic tree describing the grouping of populations. For example, all specimens studied in the paper belong to the *Carpiodes* genus, which as currently recognized, has three species complexes, *C. carpio*, *C. cyprinus*, and *C. velifer*. If we view the *Carpiodes* genus as the parent node, it has three leaf nodes (at the same level) corresponding to *C. carpio*, *C. cyprinus*, and *C. velifer*, respectively. This is the taxonomic tree for *Carpiodes*, which is known in biology. The classification tree in Figure 5 is a binary tree because all classifiers are binary. From an input-output point of view, the binary classification tree implements the taxonomic tree for *Carpiodes*. The structure of the classification tree is a design issue independent to the taxonomic tree.

For a given collection of samples $\mathbf{x}_i$ with the corresponding labels $y_i \in \{-1, 1\}$, designing a binary classifier can be solved by any conventional supervised learning algorithm. However, we argue that feature selection is indispensable in our system for the following reasons. From a taxonomic viewpoint, it is desirable to use a small number of body shape characters to uniquely diagnose a species as distinct

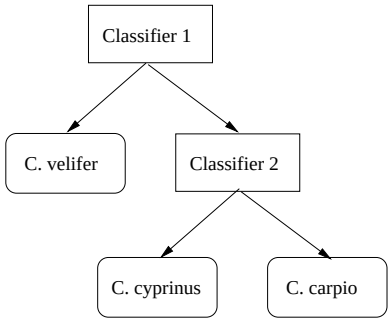


Figure 5: A hierarchical classifier for *Carpiodes* genus.

from its known relatives. The feature selection procedure can identify those "most" diagnostic features (in this case, body shape characters). From a machine learning viewpoint, constraining the number of selected features is an effective way to avoid overfitting. The experimental results in Section 6.4 also demonstrate the efficacy of feature selection in avoiding potential overfitting.

Feature subset selection is a well-researched topic in the areas of statistics, machine learning, and pattern recognition [15, 30]. Existing selection approaches generally fall into two categories: filter and wrapper [15, 30]. Some filter methods such as ranking through correlation coefficients or through Fisher scores tend to select inter-correlated features and does not guarantee an acquisition of a good classifier. On the contrary, wrappers include the desired classifier as a part of their performance evaluation, which is a joint feature selection and classification approach. They tend to produce better generalization but may require expensive computational cost.

The proposed approach is a wrapper model based on 1-norm SVM. Consider the problem of finding a linear classifier

$$y = \text{sign}\left(\mathbf{w}^T\mathbf{x} + b\right)$$

where $\mathbf{w}$ and b are model parameters. The SVM approach constructs classifiers based on hyperplanes by minimizing a regularized training error $\lambda R[\cdot] + error$ where $R[\cdot]$ is a regularization operator, λ is called the regularization parameter, and $error$ is commonly defined through a hinge loss function

$$\xi = \max\{1 - y(\mathbf{w}^T\mathbf{x} + b), 0\} .$$

When an optimal solution $\mathbf{w}$ is obtained, the magnitude of its component w_k indicates the significance of the effect of the k-th feature on the classifier. Those features corresponding to a non-zero w_k are selected and used in the classifier.

The regularization operator in standard SVMs is the 2-norm of the weight vector $\mathbf{w}$, which formulates SVMs as quadratic programs (QP). Solving QPs is typically computationally more expensive than solving linear programs (LPs). SVMs can be transformed into LPs as in [33]. This is achieved by regularizing with a sparse-favoring norm, i.e., the 1-norm of $\mathbf{w}$, $\|\mathbf{w}\|_1 = \sum_k |w_k|$. Thus 1-norm SVM is also referred to as sparse SVM and has been similarly applied to other practical problems such as drug discovery in [3].

Many practical problems in image classification relate to imbalances in samples, i.e., the number of negative samples is much larger than the number of positive samples. To tackle this imbalanced issue and make classifiers biased towards the minority class, we penalize differently on errors produced respectively by positive samples and by negative ones. Rewrite $w_k = u_k - v_k$ where $u_k, v_k \geq 0$. If either u_k or v_k has to equal to 0, then $|w_k| = u_k + v_k$. The LP is formulated in variables $\boldsymbol{\theta} = \{\mathbf{u}, \mathbf{v}, b, \boldsymbol{\xi}, \boldsymbol{\eta}\}$ as

$$\begin{array}{ll} \min_{\boldsymbol{\theta}} & \lambda \sum_{k=1}^{d} (u_k + v_k) + \frac{\mu}{\ell^+} \sum_{i=1}^{\ell^+} \xi_i + \frac{1-\mu}{\ell^-} \sum_{j=1}^{\ell^-} \eta_j \\ \text{s.t.} & \left[(\mathbf{u} - \mathbf{v})^T \mathbf{x}_i^+ + b\right] + \xi_i \geq 1, i = 1, \cdots, \ell^+, \\ & -\left[(\mathbf{u} - \mathbf{v})^T \mathbf{x}_j^- + b\right] + \eta_j \geq 1, j = 1, \cdots, \ell^-, \\ & u_k, v_k \geq 0, k = 1, \cdots, d, \\ & \xi_i, \eta_j \geq 0, i = 1, \cdots, \ell^+, j = 1, \cdots, \ell^-. \end{array}$$

where $\mathbf{x}_i^+$ and $\mathbf{x}_j^-$ denote a positive sample and a negative sample, respectively, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are hinge losses, $0 < \mu < 1$ is a constant penalizing the errors from positive and negative samples, ℓ^+ (ℓ^-) is the number of positive (negative) samples.

5. SIMILARITY MATCHING

The image similarity measure consists of two parts. The first part corresponds to the semantic similarity, which is determined by semantic classifier in Section 4. If two specimens belong to the same semantic class, their similarity is the maximum, otherwise the similarity is zero. Specifically, the semantic similarity between two specimens $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as

$$s_1(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1 & \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are in the same class} \\ 0 & \text{otherwise} \end{cases}$$

The second part reflects the overall shape similarity, and is defined as

$$s_2(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}}$$

where σ^2 is chosen to be the sample variance of the overall shape distance. Note that $\|\mathbf{x}_i - \mathbf{x}_j\|$ is the distance in the shape space, hence describes the overall shape difference between two specimens. The similarity measure is then defined as a convex combination of semantic similarity and overall shape similarity:

$$s(\mathbf{x}_i, \mathbf{x}_j) = \alpha\, s_1(\mathbf{x}_i, \mathbf{x}_j) + (1 - \alpha)\, s_2(\mathbf{x}_i, \mathbf{x}_j) \qquad (1)$$

where $\alpha \in [0, 1]$ is a parameter specified by a user.

Note that labels are included for all images in the database. This is in contrast to a typical CBIR system where the labels (or semantic annotations) for the images in the database are usually not available. A small distance in overall shape does not necessarily imply semantic similarity because the semantic classification is based on a small number of selected shape characters rather than the overall shape. The above similarity measure gives users flexibility in retrieval: one can retrieve specimens based on the overall shape similarity ($\alpha = 0$), predicted species label ($\alpha = 1$), or a convex combination of the two ($0 < \alpha < 1$).

Although the ranking of images are based on the landmarks (label of the query is computed from the landmark coordinates), we argue that the images are indispensable because: (1) It is much easier to visualize the specimens using images than using landmarks alone; (2) The system might misdiagnose the query. It is easier for taxonomists to identify the errors based on the retrieved images than on the coordinates of the landmarks.

6. EXPERIMENTAL RESULTS

We test the proposed CBIR system on the specimens from the three *Carpiodes* morphotypes: *C. carpio*, *C. cyprinus*, and *C. velifer*. The current database contains only 600 images of *Carpiodes* specimens. However, the proposed computational framework can be applied to any number of images at any level of fish taxonomy. We are working to expand the database by including images of specimens of a related group of suckers in the genus *Ictiobus*. Our experiments consists of two steps:

- Demonstrating the efficacy of semantic classification by identifying features (or body characters) for distinguishing among the three *Carpiodes* morphotypes;
- Applying the system to a taxonomic revision problem involving populations from Colorado River in Texas and Rio Grande and comparing the results with those based on the DNA analysis.

6.1 System Interface

The system has a simple CGI-based query interface. Users can either enter the ID of an image as the query or submit any image (along with a file containing the landmarks) via the Internet. Figure 6 shows the 25 thumbnails returned by the system where the query image (*C. Cyprinus*) is on the top left. The parameter α in (1) was chosen to be 0.8. Below each thumbnail are image ID and the name of its taxonomic category. Users can start a new query search by submitting a new image ID or image files.

6.2 Semantic Classification

We apply the 1-norm SVM to select features and build classifiers simultaneously. The binary classifiers are organized as in Figure 5. Two parameters, λ and μ, need to be specified for 1-norm SVM. We set μ to be the percentage of negative training samples to balance the training errors on positive and negative samples. The regularization parameter λ is selected such that at most three features are

Figure 6: The interface of the CBIR system.

Table 2: Semantic classification of *Carpiodes* into three morphotypes: *C. carpio*, *C. cyprinus*, and *C. velifer*. The hierarchical classifier first separates *C. velifer* from the rest species. It then distinguishes *C. carpio* and *C. cyprinus*.

Classification Problem	Selected Features	Training Error	Test Error
C. velifer/the rest	x_{10}, x_{11}	10%	11.7%
C. carpio/*C. cyprinus*	x_4, x_7	12.9%	13.9%

selected. This is based upon the fact that taxonomists rarely use more than three body shape characters to describe the difference among closely related species or subspecies.

The images within each class are randomly divided into a training set and a test set of equal size. The classification results are summarized in Table 2. The hierarchical classifier first separates *C. velifer*-like specimens from specimens of other species. It then distinguishes *C. carpio* from *C. cyprinus*. In all the experiments, we observe that the performance based on three selected features is similar to that based on two selected features. The selected features in both classification problems are also shown in Table 2.

6.3 Species Prediction

The experiment is based on 53 specimens from upper Colorado River in Texas and the Rio Grande. They were traditionally recognized as *C. carpio*, yet recent DNA evidence suggests that both populations are more closely related to *C. cyprinus*. So we view these 53 specimens as "suspicious" populations. Each "suspicious" specimen is submitted to the system as a query image. The predicted class label of the query is determined by the majority class among the top k retrieved images (specimens). We observed that the results are robust for k varying from 10 to 60. So we pick $k = 20$.

We first set the parameter α in the similarity measure (1) 0.1. This corresponds to retrieving specimens that are similar to the query based on the overall shape. It turns out that 52 out of the 53 suspicious specimens are recognized as *C. carpio*, and only 1 specimen is identified as *C. cyprinus*. In other words, the overall shape suggests that the "suspicious" specimens should be classified as *C. carpio*. Next, we increase α to 0.9, i.e., the decision is based mainly on the semantic classifiers designed in Section 6.2. In this case, 23 "suspicious" specimens are classified as *C. carpio*, while the remaining 30 specimens are classified as *C. cyprinus*. We get identical results for $\alpha = 1.0$.

Although the hierarchical classifier in Section 6.2 can distinguish the three species with reasonable accuracy using only four body shape characters, it has difficulty categorizing the specimens from Colorado River in Texas and Rio Grande as either *C. carpio* or *C. cyprinus*; 43.4% of the "suspicious" specimens are assigned to *C. carpio*, and 56.6% to *C. cyprinus*. At the same time, the retrieved images based on overall shape identify 52 out of 53 specimens as *C. carpio*. These contradictory results can be viewed as an indication that the suspicious specimens represent a new species. It is very interesting that overall shape analysis and the DNA analysis give similar results: the suspicious specimens are more similar to *C. carpio* than to *C. cyprinus* in terms of the overall shape, yet they are genetically closer to *C. cyprinus*. Note that our CBIR system can easily obtain a similar

Table 3: Semantic classification of *Carpiodes* using all 12 features. The hierarchical classifier first separates *C. velifer* from the rest species. It then distinguishes *C. carpio* and *C. cyprinus*.

Classification Problem	Classifier	Training Error	Test Error
C. velifer versus the rest	SVM (linear)	9.1%	9.5%
	SVM (Gaussian)	8.9%	9.8%
C. carpio versus *C. cyprinus*	SVM (linear)	16.9%	17.5%
	SVM (Gaussian)	16.5%	17.3%

conclusion by adjusting the value of the parameter α.

6.4 Discussions

We did experiments to see whether feature selection is indispensable in semantic classification. The semantic classification results using all 12 features are shown in Table 3. We tested two classifiers, namely, linear SVM and SVM with Gaussian kernel. All the classifiers were constructed using half of the 600 specimens and tested over the remaining 300 specimens. As Table 3 indicates, the new classification results are similar to those in Table 2, i.e., with feature selection. However, the new classifiers generate significantly different predictions on the 53 "suspicious" specimens. Both classifiers assign the majority of the 53 specimens to *C. carpio*, which contradicts the results generated by 1-norm SVM.

An interesting question arises: which results should we trust, those based on the selected features or those using all the features? We argue that feature selection is indispensable for the following reasons:

- From a taxonomic viewpoint, it is desirable to use a small number of body shape characters to describe a species as distinct from its known relatives. The feature selection procedure can identify those "most" diagnostic features (or body shape characters).
- From a machine learning viewpoint, constraining the number of selected features is an effective way to avoid overfitting. One may reason that the above conflicting result for Colorado River and Rio Grande specimens is due to overfitting, i.e., the models trained on all 21 features overfit the data.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a content-based image retrieval approach for taxonomic research. The system has a learning component that automatically identifies the semantic class of a query based on digitized landmarks. We applied the system to a taxonomic problem in genus *Carpiodes*. The results are promising: the proposed framework not only learned classifiers that well separated the three known species in *Carpiodes* using only a few body shape features, but also recognized "suspicious" specimens that could not be identified previously without the aid of DNA analysis. Therefore, our framework provides a powerful tool for assisting the diagnosis of new species and increasing the pace of taxonomic research.

As continuations of this work, several directions may be pursued. Our system can be linked to the Internet so that taxonomists around the globe can not only retrieve specimens from the system, but can contribute images to expand the database. The learning component in the system can potentially be extended to any taxonomic problem involving a large data set and a significant percentage of unknown specimens in a semi-supervised learning framework. An important future direction of this research is to automatically build a classification tree of recognized taxa (species). Classification using shape contexts, as in the approach proposed by Belongie et al. [2], is another interesting direction to investigate.

8. ACKNOWLEDGMENTS

This work was supported in part by grants from the Louisiana Board of Regents (RCS and EPSCoR PFUND), US National Science Foundation (DEB-0237013 to HLB), The Research Institute for Children, and University of New Orleans. The authors would like to thank Jinbo Bi, Huimin Chen, and Shuqing Huang for helpful discussions.

9. REFERENCES

[1] D. C. Adams, F. J. Rohlf, and D. E. Slice. Geometric morphometrics: ten years of progress following the 'revolution'. *Ital. J. Zool.*, 71:5-16, 2004.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(4):509-522, 2002.

[3] J. Bi, K. P. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3:1229-1243, 2003.

[4] F. L. Bookstein. *Morphometric Tools for Landmark Data*. Cambridge University Press: Cambridge, 1991.

[5] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(8):1026-1038, 2002.

[6] Y. Chen and J. Z. Wang. A region-based fuzzy feature matching approach to content-based image retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(9):1252-1267, 2002.

[7] 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*, 9(1):20-37, 2000.

[8] A. Del Bimbo and P. Pala. Visual image retrieval by elastic matching of user sketches. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(2):121-132, 1997.

[9] M. Ferecatu, M. Crucianu, and N. Boujemaa. Sample selection strategies for relevance feedback in region-based image retrieval. *Lecture Notes in Computer Science*, 3332:497-504, 2004.

[10] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 3(3-4):231-262, 1994.

[11] T. Gevers and A. W. M. Smeulders. PicToSeek: combining color and shape invariant features for image retrieval. *IEEE Trans. Image Processing*, 9(1):102-119, 2000.

[12] A. Gupta and R. Jain. Visual information retrieval. *Communications of the ACM*, 40(5):70-79, 1997.

[13] *International Code of Zoological Nomenclature, 4th edition*. International Trust for Zoological Nomenclature, c/o Natural History Museum, 1999.

[14] D. G. Kendall. Shape-manifolds, procrustean metrics and complex projective spaces. *Bulletin of the London Mathematical Society*, 16:81-121, 1984.

[15] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273-324, 1997.

[16] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(9):1075-1088, 2003.

[17] W. Y. Ma and B. Manjunath. NeTra: a toolbox for navigating large image databases. In *Proc. IEEE Int'l Conf. on Image Processing*, pages 568-571. 1997.

[18] S. Mehrotra, Y. Rui, M. Ortega-Binderberer, and T. S. Huang. Supporting content-based queries over images in MARS. In *Proc. IEEE Int'l Conf. on Multimedia Computing and Systems*, pages 632-633. 1997.

[19] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: content-based manipulation for image databases. *International Journal of Computer Vision*, 18(3):233-254, 1996.

[20] S. L. Pimm and J. H. Lawton. Ecology–planning for biodiversity. *Science*, 279:2068-2069, 1998.

[21] J. E. Rodman and J. H. Cody. The taxonomic impediment overcome: NSF's partnerships for enhancing expertise in taxonomy (PEET) as a model. *Systematic Biology*, 52:428-435, 2003.

[22] F. J. Rohlf and F. L. Bookstein. *Proceedings of the Michigan Morphometrics Workshop*, No. 2. The University of Michigan Museum of Zoology, 1990.

[23] S. Santini and R. Jain. Similarity measures. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(9):871-883, 1999.

[24] 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 Analysis and Machine Intelligence*, 22(12):1349-1380, 2000.

[25] J. R. Smith S.-F. Chang. VisualSEEK: a fully automated content-based query system. In *Proc. 4th ACM Int'l Conf. on Multimedia*, pages 87-98. 1996.

[26] R. D. Suttkus and H. L. Bart, Jr. A preliminary analysis of the river carpsucker, Carpiodes Carpio, in the southern portion of its range. In *L. Lozano (ed.) Libro jubilar en honor al Dr. Salvador Contreras Balderas*, Universidad Autonoma de Nuevo Leon, Monterrey Mexico, pages 209-221. 2002.

[27] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proc. 9th ACM Int'l Conf. on Multimedia*, pages 107-118. 2001.

[28] J. Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(9):947-963, 2001.

[29] Q. D. Wheeler, P. H. Raven, and E. O. Wilson. Taxonomy: impediment or expedient? *Science*, 303:285, 2004.

[30] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy, *Journal of Machine Learning Research*, 5:1205-1224, 2004.

[31] M. Zelditch, D. Swiderski, D. Sheets, and W. Fink. *Geometric Morphometrics for Biologists: a Primer*. Elsevier Academic Press: London, 2004.

[32] X. S. Zhou and T. S. Huang. Comparing discriminating transformations and SVM for learning during multimedia retrieval. In *Proc. 9th ACM Int'l Conf. on Multimedia*, pages 137-146. 2001.

[33] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. *Advances in Neural Information Processing Systems, 16*. 2004.