Integrated Feature Selection and Clustering for Taxonomic Problems within Fish Species Complexes

Huimin Chen
Dept. of Electrical Engineering
University of New Orleans
New Orleans, LA 70148
Email: hchen2@uno.edu

Henry L. Bart, Jr.
Dept. of Ecology & Evolutionary Biology
Tulane University
New Orleans, LA 70118
Email: hank@museum.tulane.edu

Shuqing Huang
General Dynamics Information Technology
1201 Elmwood Park Blvd
New Orleans, LA 70123
Email: sophie.huang@gdit.com

***Abstract*— As computer and database technologies advance rapidly, biologists all over the world can share biologically meaningful data from images of specimens and use the data to classify the specimens taxonomically. Accurate shape analysis of a specimen from multiple views of 2D images is crucial for finding diagnostic features using geometric morphometric techniques. We propose an integrated feature selection and clustering framework that automatically identifies a set of feature variables to group specimens into a binary cluster tree. The candidate features are generated from reconstructed 3D shape and local saliency characteristics from 2D images of the specimens. A Gaussian mixture model is used to estimate the significance value of each feature and control the false discovery rate in the feature selection process so that the clustering algorithm can efficiently partition the specimen samples into clusters that may correspond to different species. The experiments on a taxonomic problem involving species of suckers in the genus *Carpiodes* demonstrate promising results using the proposed framework with only a small size of samples.**

***Index Terms*— feature selection, clustering, taxonomy, shape analysis, false discovery rate, image fusion**

I. Introduction

Biologists have traditionally consulted field guides and other published works to identify species that they encounter in the field and to summarize what is known about the biology of those species. However, these guides rarely contain complete information on species identity, distribution and biology. Much of this information resides with specimens in natural history museums, inaccessible to most biologists. Existing information systems of natural history museums are mainly taxonomically focused. They are designed to give the research community global access to specimen information for various named species or higher taxonomic groups. However, the names assigned to specimens are not always the most up-to-date, or the specimens may belong to groups that have not been studied and fully resolved taxonomically.

The job of identifying and describing new species and determining interrelationships among species falls on taxonomists and systematists. Taxonomy and systematics, as traditionally practiced, can be painfully slow. The reason for this is that taxonomists typically have to examine and gather data from large numbers of specimens across broad geographical areas in order to identify the features that uniquely diagnose a new species from related known species. As a consequence, it is estimated that only 10% of the world's species have been discovered and described. The pace of new species discovery and description would speed up significantly if multimedia and machine learning techniques could be developed to automatically identify diagnostic features of specimens archived in natural history museums.

Geometric morphometrics [20], as a well developed technique, has been widely used in diagnosing fish species [1]–[3]. The idea is to use landmarks to characterize shape variation among the specimens of different species. Computer-based statistical methods such as multivariate analysis [4] are often applied to various taxonomic problems with many successful stories [23]. However, understanding why and how morphological differences have emerged is challenging since body shape has a genetic basis but is also subject to epigenetic and environmental processes. An alternative is to apply outline analysis [11] or eigenshape analysis [12] where more information than the homologous landmarks is explored to derive biologically meaningful features. As the advances of efficient machine learning and data mining algorithms [13], a new computational framework has been developed [8] to jointly select features and classify fish species. One interesting question is whether a good clustering algorithm can automatically select useful features to quantitatively compare the similarity among specimens.

Feature selection algorithms for clustering largely fall into three categories: the filter model [10], the wrapper model [6], and the hybrid model [9]. The filter model

This paper is based on "Integrated Feature Selection and Clustering from Multiple Views for a Taxonomic Problem," by H. Chen, H. L. Bart, and S. Huang, which appeared in the Proceedings IEEE 9th Workshop on Multimedia Signal Processing (MMSP 2007), Chania, Crete, Greece, October 2007, © 2007 IEEE.

This work was supported in part by Air Force Research Lab # FA8650-07-M-1161 and Navy Air through Planning Systems Inc. Contract # N68335-05-C-0382.

relies on general characteristics of the data to select the feature subset which are hard to obtain from the taxonomists. The wrapper model uses a prespecified clustering algorithm to judge the relevance of the feature subset. It relies critically on the clustering algorithm and tends to be computationally expensive. The hybrid model uses a filter based criterion to direct the search over feature subsets but makes wrapper based feature selection.

The proposed feature selection framework is closely related to the hybrid model. We fit each feature variable by a Gaussian mixture and select the feature subset by controlling the false discovery rate [5]. Our approach is computationally efficient and can construct a binary cluster tree for taxonomists to perform further analysis. Two types of features are generated as the candidates. The shape induced features from reconstructed 3D shape are generated from multiple views of each specimen. Local saliency characteristics are generated directly from the 2D images. In our study of fish genus *Carpiodes*, we found that one species, namely *C. velifer* separates well from the other two, *C. carpio* and *C. cyprinus*. However, *C. carpio* and *C. cyprinus* do not form as well separated clusters. Most of the specimens from the Rio Grande and upper Colorado River in Texas, which are presently classified taxonomically as *C. carpio*, fall into the *C. cyprinus*-like cluster, a result in agreement with recent DNA sequencing results.

In summary, our contributions include: (1) defining a framework for joint feature selection and clustering with possibly correlated feature set which is generally applicable to many taxonomic problems; (2) demonstrating that our framework is effective in constructing a binary cluster tree within fish species complexes using shape and saliency features from specimen images taken in three different views; and (3) revealing that our framework can help taxonomists expedite the diagnosis of specimens for the revision of existing taxonomy and the discovery of new species.

The rest of the paper is organized as follows. Section II describes a challenging taxonomic problem involving specimens from the genus *Carpiodes* where geometric morphometrics can lead to controversial taxonomic results. Section III presents the computational framework for integrated feature selection and clustering. Section IV discusses the feature generation method and feature selection criterion. Section V presents the experimental results on the taxonomic problem using the proposed feature selection and clustering method. Concluding remarks are given in Section VI.

II. Motivating Example

We start with a discussion of a taxonomic problem involving suckers of genus *Carpiodes*. 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*). Most taxonomists regard each of these species as

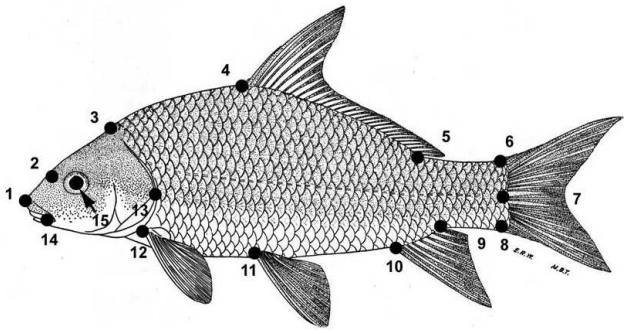


Figure 1. Digitized 15 homologous landmarks using TpsDIG Version 1.4 (by F. J. Rohlf).

complexes of multiple species in need of revision [21]. The goal of *taxonomic revision* in this case is to identify and formally describe the unrecognized species.

Geometric morphometrics has been applied to analyze the variation in body shape using a collection of biologically definable landmarks (also called homologous landmarks) along the body [2], [3]. Capturing geometry by a way of landmark data has become rather commonplace. Landmarks are precise locations on biological forms that hold some developmental, functional, structural, or evolutionary significance. Figure 1 shows 15 homologous landmarks digitized on a specimen using TpsDIG software tool developed by F. J. Rohlf of SUNY Stony Brook (`http://life.bio.sunysb.edu/morph/`). The analysis methods accompanying the software focus on the coordinates of landmarks and the 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 the overall shape space. Unfortunately, the current geometric morphometric methods have two major limitations:

- Groups of specimens are distinguished from other populations based on a small set of derived variables, which are usually functions, in the 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 used in defining new species, and thus cannot be formally used to describe the unrecognized species.
- Shape variation of specimens from closely related species or subspecies may not be discernible in the overall shape space or using the analysis based on landmark coordinates. Therefore, existing geometric morphometric methods may generate misleading results.

Over the years since [21] was published, H. L. Bart has examined shape and DNA sequence variation in many *Carpiodes* populations. Figure 2 shows the results of an analysis of overall body shape based on a geometric morphometrics using canonical variate analysis (CVA). CVA grouped specimens from the Rio Grande (squares), upper

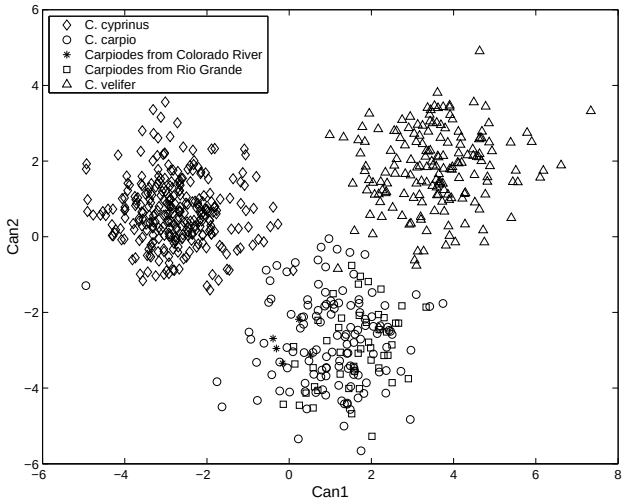


Figure 2. Plot of 650 *Carpiodes* specimens representing three distinct morphotypes on the first two canonical variate axes based on derived shape variables from geometric morphometric analysis of landmark data.

Colorado River (stars), and other western Gulf Slope rivers cluster with *C. carpio* specimens (circles) from the Mississippi River Basin. However, a surprising discovery 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 H. L. Bart three years of careful study among 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 attempt to address next is: Can computer aided feature selection and clustering techniques be applied to diagnose taxonomic groups in genus *Carpiodes* more accurately than geometric morphometrics?

III. Integrated Feature Selection and Clustering

We first present the theoretical formulation of a general taxonomic problem using mixture model with unknown parameters. We then discuss the feature selection and its impact on the clustering result. To avoid exhaustive enumeration over all feasible feature subsets, we propose to use an efficient step-down testing procedure by controlling the false discovery rate.

A. Mixture Based Clustering

To construct a binary cluster tree, we denote N input samples by $\mathbf{Y}_N = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$. Assume that each sample $\mathbf{y}_i$ can either come from group 1 with probability α_1, which has the likelihood function $p(\mathbf{y}_i|\theta_1)$, or from group 2 with probability α_2, which has the likelihood function $p(\mathbf{y}_i|\theta_2)$. Thus the log-likelihood of two-component mixture can be written as

$$\log p(\mathbf{Y}_N|\Theta) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{2} \alpha_j p(\mathbf{y}_i|\theta_j) \right)$$

where $\{\alpha_j\}$ satisfies $\alpha_1 > 0$; $\alpha_2 > 0$; $\alpha_1 + \alpha_2 = 1$. The unknown parameter to be estimated from the input samples is denoted by $\Theta = \{\theta_1, \theta_2, \alpha_1, \alpha_2\}$. The maximum likelihood estimate $\hat{\Theta}^{\mathrm{ML}}$ can be obtained using the following iterative procedure.

1) Make an initial guess Θ^{old}.
2) Compute the mixture probabilities for each sample as follows.

$$\beta_{ji} = \frac{\alpha_j^{\mathrm{old}} p(\mathbf{y}_i|\theta_j^{\mathrm{old}})}{\sum_{j=1}^{2} \alpha_j^{\mathrm{old}} p(\mathbf{y}_i|\theta_j^{\mathrm{old}})}.$$

3) Update the mixture probability

$$\alpha_j^{\mathrm{new}} = \frac{1}{N} \sum_{i=1}^{N} \beta_{ji}.$$

4) Identify the indices of input samples which clusters to group j

$$\mathcal{S}_j = \{i | \beta_{ji} > 0.5\}.$$

5) Update the parameter estimate for each component

$$\theta_j^{\mathrm{new}} = \arg\max_{\theta_j} \sum_{i \in \mathcal{S}_j} \log p(\mathbf{y}_i|\theta_j).$$

6) Update with $\Theta^{\mathrm{old}} = \Theta^{\mathrm{new}}$. Repeat steps 2)–5) until Θ^{new} converges.

The algorithm described above can be seen as a special type of the expectation maximization (EM) procedure [18]. It guarantees to converge to a stationary point of the likelihood function, but not necessarily the global maximum. In practice, one should examine whether $|\mathcal{I}_1|$ and $|\mathcal{I}_2|$ are well balanced and make a few random initial guesses to avoid being trapped in a local maximum.

Based on the clustering results, one can continue the procedure to each subset of the samples until all input samples are partitioned into a binary cluster tree. Note that the parameters θ_1 and θ_2 can reside in different spaces or have different dimensions. The reliability of each binary clustering step depends on the separability between $p(\mathbf{y}|\theta_1)$ and $p(\mathbf{y}|\theta_2)$, which can be measured using K-L divergence [18].

B. Feature Selection in Mixture Based Clustering

Assume that each sample $\mathbf{y}_i = [y_{i1}^T \ y_{i2}^T \ ... \ y_{id}^T]^T$ contains d candidate features and only a small subset of the features is relevant to the mixture based clustering. Denote by $\mathbf{Y}_{N(\mathcal{I}_K)}$ the input samples with K features indexed by $\mathcal{I}_K$ being selected from the d candidates. With the conditional independence assumption among the selected features, the log-likelihood of two-component mixture can be written as

$$\log p(\mathbf{Y}_{N(\mathcal{I}_K)}|\Theta_{(K)}) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{2} \alpha_j \prod_{l=1}^{K} p(y_{i(l)}|\theta_{j(l)}) \right).$$

Note that the unknown parameter for each cluster depends on the selected feature subset. For example, the input sample is a 12 dimensional vector but only the first and second component are from a Gaussian mixture distribution while the rest components are from a single Gaussian distribution. In this case, $d = 12, K = 2$ and the best feature subset is $\mathcal{I}_k = \{1, 2\}$. Selecting any other feature will deteriorate the clustering performance.

Denote by $H_j(\mathcal{I}_K, \Theta_{(K)})$ the hypothesis that the feature subset has the index set $\mathcal{I}_K$ with mixture parameter $\Theta_{(K)}$. It is tempting to select the hypothesis that yields the maximum likelihood, i.e.,

$$(\hat{\mathcal{I}}_K^{\mathrm{ML}}, \hat{\Theta}_{(K)}^{\mathrm{ML}}) = \arg \max_{(\mathcal{I}_K, \Theta_{(K)})} \log p(\mathbf{Y}_{N(K)}|\Theta_{(K)}).$$

However, it does not make too much sense to compare the hypotheses with the underlying likelihood functions in different parameter spaces. A more complicated parametric model can easily overfit the observation data when sample size is small. Thus one has to introduce appropriate penalty to the log-likelihood function such as using the minimum description length (MDL) criterion [19]. Note that the number of hypotheses grows exponentially in d which makes the feature subset selection very inefficient. Next, we propose a less accurate but more efficient feature selection method that only controls the false selection rate to be below a desired level.

C. Efficient Feature Selection and Clustering by Controlling False Discovery Rate

To avoid exhaustive search over all feature subsets, we treat feature selection problem as multiple hypothesis testing with the following problem formulation. A hypothesis $H_j(\mathcal{I}_K)$ describes the index set $\mathcal{I}_K \subseteq \{1, \cdots, d\}$ of the input sample, i.e., the selected feature subset. Formally, we can write the hypothesis
$H_j(\mathcal{I}_K)$: $\theta_{1(i)} \neq \theta_{2(i)}$ if $i \in \mathcal{I}_K$, otherwise $\theta_{1(i)} = \theta_{2(i)}$.

In the above formulation, the input samples from the selected feature subset are from a mixture distribution while the rest of the input variables are assumed to be from a single parametric distribution. This makes all hypotheses have the same parameter space. The selection of feature subset can be viewed as testing d hypotheses ($\theta_{1(i)} = \theta_{2(i)}$ vs. $\theta_{1(i)} \neq \theta_{2(i)}$, $i = 1, ..., d$) simultaneously. Among those features that are truly from a mixture distribution, we want to control the percentage of selecting non-diagnostic features from the d candidates. The control of false discovery rate (FDR) is less strict than the control of family-wise error rate but more reasonable when d is large [5]. We want to select the feature subset for mixture based clustering with controlled FDR so that the probability of selecting diagnostic features will be higher than that using the family-wise error rate control. The proposed procedure to choose a hypothesis H_j within the desired FDR level does not require any independence assumption of the test statistics. It is a step-down test which is more efficient than the commonly used step-up FDR test [5] when the number of selected features is relatively small compared with d.

The procedure starts with the test statistic $T_1, \cdots, T_d$ based on the element-wise maximum likelihood estimate $\hat{\theta}_{(1)}, \cdots, \hat{\theta}_{(d)}$. Each test statistic T_i is associated with a p-value [18], π_i, indicating the probability of obtaining the test statistic at least as extreme as T_i when $\theta_{1(i)} = \theta_{2(i)} = \hat{\theta}_{(i)}$, i.e., the two mixtures coincide. For any user specified significance level $q \in (0, 1)$, the feature subset is selected by performing the following steps which controls the FDR to be below q.

- Order the p-values such that $\pi_{(1)} \leq \cdots \leq \pi_{(d)}$.
- Compute the index $u_i = \min\left(1, \frac{d}{(d-i+1)^2} q\right), i = 1, ..., d$.
- Reject all hypotheses $\theta_{1(j)} = \theta_{2(j)}$ for $1 \leq j \leq K-1$ where K is the smallest index for which $\pi_{(K)} > u_K$. If no such K exists, then the clustering algorithm stops.

Once the subset $\hat{\mathcal{I}}_K$ is determined, the mixture based clustering result should be recomputed using only the selected input features. Clearly, the FDR controlled feature selection procedure is much more efficient than finding the optimal feature subset via exhaustive search. The proof that the above procedure yields the FDR below the significance level q can be found in the appendix.

Note that the mixture based clustering with feature selection algorithm only needs to find the maximum likelihood estimates of d different mixture models. The associated p-values rely on the clustering algorithm, which is similar to the wrapper model.

IV. Feature Generation

A. 3D Shape Construction from Multiple Views

Three-dimensional reconstruction of objects from two-dimensional images is a fairly well explored field. The most common approach has been reconstruction of the 3D shape from multiple 2D slices, typically obtained using Computed Tomography (CT), Magnetic Resonance Imaging (MRI), etc. However, these methods are expensive and time consuming when being applied to digitize the specimens. A less expensive method is considered as follows. Given a set images of a specimen from multiple views, we want to construct the 3D locations of all landmarks on the images. Body shape and form characters can be deduced from the landmarks [20]. Figure 3 shows

13 manually marked landmarks and certain distances (in pixels) between two landmarks on a *Carpiodes* specimen from upper Colorado River. Its species is yet to be determined.

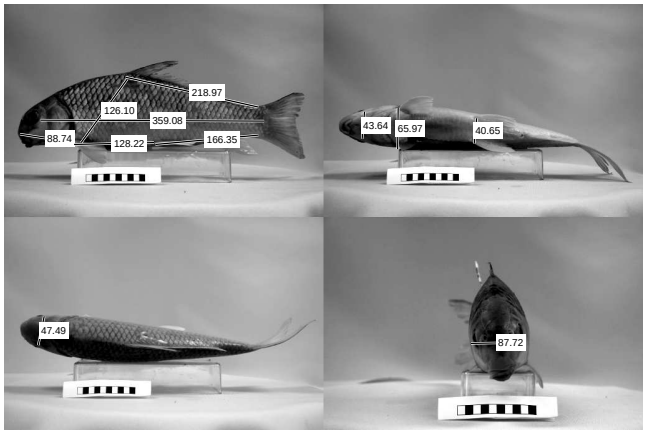


Figure 3. Multiple views of a specimen with landmarks from upper Colorado River in Texas.

From a mathematical point of view, morphology is a set theoretic method to image processing characterized by selective filtering of data at every stage, so that only desired artifacts can be isolated and later recognized. Here we describe a few morphological operations being used to isolating and identify feature regions.

- Thresholding: All vertices for which the value of feature-marker fitting lies within the predetermined lower and upper limit are marked as *feature* and others as *background*.
- Neighborhood: For each vertex, the set of its immediately connected neighbors is computed. The radius of a neighborhood can be recursively enlarged by increasing the depth of the connection.
- Dilation: Every vertex is examined and marked by 1 if at least one of its neighbors is marked by 1.
- Erosion: Every vertex is examined and marked by 0 if at least one of its neighbors is marked by 0.

Four paired landmarks are identified from four different views of the specimen image, namely, the two eyes, the eye and the middle point of tail, the upper fin and the upper point of tail, the lower lip and first lower fin. We use four midpoints of the four paired landmarks and the tip of head to register the coordinate system of the specimen in 3D. We adopt the parallax based approach [14] to construct the 3D shape using the landmarks and generate candidate features by removing non-shape related variations as described in [8]. A more accurate method to place landmark or semi-landmark points on complex surfaces for the purpose of registration, alignment and morphing has been developed in [22].

B. Saliency Based Features

Feature generation using landmark data alone is limited. For example, when landmark data are collected, no verifiable information regarding the surfaces that lie between the landmarks is retrievable from analysis of the data. Other salient features are also useful for discriminating different species especially when the overall body shape characteristics are similar among certain species. Gradient magnitude is traditionally used to discriminate between salient and non-salient edges but it often suffers from noise and minor variation in intensity values. Alternatively, we focus on information theoretic measure with respect to spatial locations and scales of objects in an image. Consider a grey scale image with a location x around which its circular neighborhood D is specified with adjustable radius s. The saliency at x with scale s is measured by the entropy

$$\mathcal{H}_D(s,x) = -\sum_{d\in D} p_{s,x}(d)\log p_{s,x}(d).$$

The entropy-scale characteristics of a particular neighborhood represents the local image structure. The scale at which the entropy reaches maximum is considered considered the appropriate scale for feature extraction since it is the scale at which the image becomes unpredictable or difficult to model. A inter-scale saliency criterion was proposed in [15], which selects scale s_p such that

$$s_p = \arg\max_s \mathcal{H}_D(s,x)\mathcal{W}_D(s,x)$$

where

$$\mathcal{W}_D(s,x) = \frac{s^2}{2s-1}\sum_{d\in D}\left|p_{s,x}(d) - p_{s-1,x}(d)\right|.$$

The probability density function $p_{s,x}(d)$ is estimated within the neighborhood area using the histogram. Note that the entropy measure between two adjacent scales does not always depict the interesting feature of an object accurately. A modified criterion, motivated by [17] to include both spatial and temporal saliency, was also given in [15], which selects s_p such that

$$s_p = \arg\max_s \mathcal{S}_D(s,x)$$

where

$$\mathcal{S}_D(s,x) = \mathcal{H}_D(s,x)\mathcal{W}_D(s,x)\mathcal{W}_D(s+1,x).$$

In practice, regions with saliency values $\mathcal{S}_D(s,x)$ greater than a certain threshold are selected as the candidate features. A feature is defined by its saliency region parameterized by the center x and radius s_p.

The salient regions are computed for each 2D image using the technique developed in [7]. All salient regions being detected for each specimen sample are normalized and aligned as the candidate features. Note that these features are automatically generated without the need of landmarks and very effective in separating the specimen from the background.

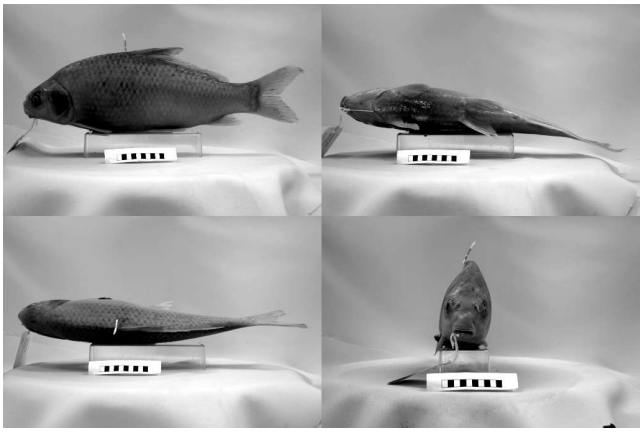
V. Experiments

A. The Taxonomic Problem

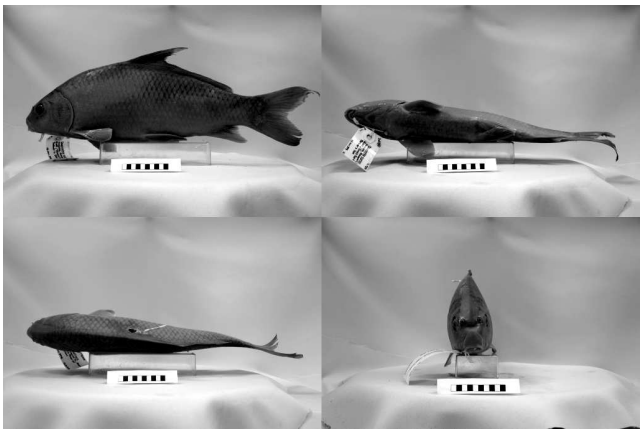
We have successfully used computer-based shape analysis methods to characterize variation in body proportions among *Carpiodes* specimens [7], [8]. However,

the use of morphometric techniques alone can generate misleading results as seen in Section II. In [8], the images of 650 *Carpiodes* specimens, each with 15 landmarks, from Tulane Museum of Natural History Fish Collection were used to identify features diagnostic for classifying them into three different species. Those features were also used to classify the *Carpiodes* specimens from the Rio Grande and upper Colorado River. Over 60% of the specimens were correctly diagnosed as *C. cyprinus* based on two statistically significant feature variables (related to the distance between the naris and the tip of the snout in proportion to the distance between the naris and the eye). Here we are interested in clustering *Carpiodes* specimens into a binary cluster tree without using the opinion from a taxonomist. We evaluate the effectiveness of the feature selection and clustering accuracy by comparing the results with those using the landmark aided feature selection for classification [8] and saliency based feature selection for classification [7].

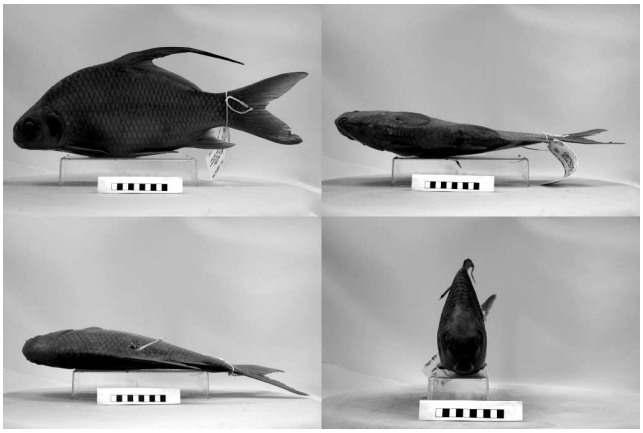
Another interesting experiment is to cluster specimens of *Carpiodes* from Rio Grande and upper Colorado River with other samples from known species. We would like to see how close they are related to *C. cyprinus* and *C. carpio*. Figure 4 shows three *Carpiodes* specimens from different species, namely, *C. carpio*, *C. cyprinus* and *C. velifer*. The side view provides most important information on the body shape while the front, top and bottom views provide detailed characteristics around the head, snout and possible lip nipple of each specimen. Note that the diagnostic features among these species are quite subtle.



(a) *C. carpio*



(b) *C. cyprinus*



(c) *C. velifer*

Figure 4. Sample images of *Carpiodes* from three different species.

B. Clustering Result

We digitized 55 specimens and took four different views (left, bottom top, front) of each specimen sample. Among those 55 samples, 26 specimens are *C. carpio* (labeled 1-26); 10 are *C. cyprinus* (labeled 27-36); 10 are *C. velifer* (labeled 37-46); 4 are from the Rio Grande (labeled 47-50) with their species undetermined; and 5 are from upper Colorado river in Texas (labeled 51-55) with their species undetermined. In the first level of the binary cluster tree, all specimens of *C. velifer* cluster into one group. The remaining specimens form another group. In the second level of the cluster tree, 3 specimens of *C. cyprinus* cluster with a group comprising 23 *C. carpio* specimens. Another cluster contains 3 specimens of *C. carpio* and 7 specimens of *C. cyprinus*. Interestingly, all of the 4 specimens from the Rio Grande cluster with the *C. cyprinus* group and 2 out of 5 specimens from the upper Colorado River in Texas also cluster with this group. This clustering result seems to be a strong indication that the unknown specimens should not be classified as *C. carpio* as traditionally held. Figure 5 shows the complete dendrogram of the clustering result. We set the false discovery rate to be below 0.01 and consider the candidate features from both normalized 3D shape characters and saliency regions. The distance of each level of the binary cluster tree is normalized using the K-L divergence between the estimated two mixture densities. We can see that the distinction between those specimens belonging to *C. carpio* and those belonging to *C. cyprinus* is not as significant as the distinction between those specimens belonging to *C. velifer* and the rest of the specimens.

We trained a logistic regression classifier with three selected features [8] that can successfully classify the 36 specimens from *C. carpio* and *C. cyprinus* without any

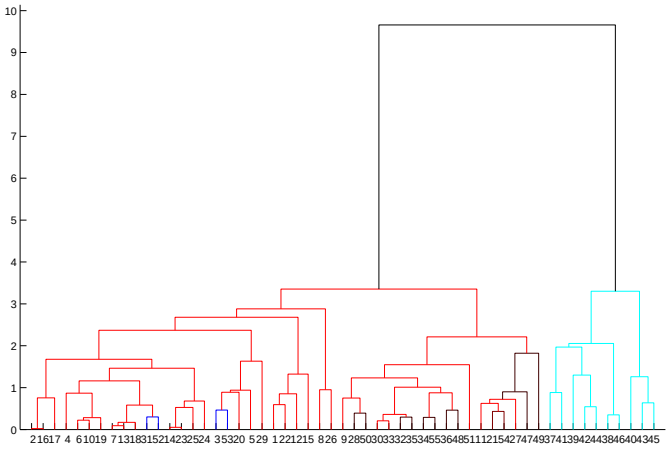


Figure 5. Dendrogram of 53 specimen samples using five selected features.

error. In the testing phase, 3 out of 4 specimens from the Rio Grande and 4 out of 5 specimens from upper Colorado river in Texas are classified as *C. cyprinus*. Thus the supervised learning confirms our clustering results. It again indicates a different conclusion from the traditional taxonomic practice and provides adequate evidence for a taxonomist to reexamine the existing species for possible revision.

C. Selected Features

At present, no method is tailored to the problem of finding diagnostic features in morphometric data. Thus selecting a small subset of feature variables which separates the specimen samples in two groups with maximal distance is crucial for a taxonomist to judge them to be different species [16]. For the first level clustering, two saliency based features were selected among 25 candidates and two 3D shaped induced features were selected among 32 candidates. For the second level clustering, three saliency based features were selected and two 3D shaped induced features were selected. The saliency based features being selected are centered around the head from left, bottom and top views, which is closely related to the diagnostic difference in snout and mouth size between *C. carpio* and *C. cyprinus*. The shape induced features being selected have strong correlation with the features selected for classification using a logistic regression classifier [8]. Thus our approach is very effective in clustering samples from *Carpiodes* into different species using an FDR controlled feature selection procedure even with small sample size.

We also tested the clustering algorithm based on the selected saliency features alone and compared with the supervised classifier with the same selected features as in [7]. For the first level clustering, the error is 12%. For the second level clustering, the error increases to 28% without considering the specimens from the Rio Grande and upper Colorado river in Texas. Thus the 3D shape induced features from the 2D images play an important role in clustering the specimens from different species.

VI. Discussion and Conclusions

We have developed an integrated feature selection and clustering framework that automatically identifies a set of diagnostic feature variables to group specimens into a binary cluster tree. The key step is to model specimens with diagnostic feature variables based on a mixture distribution while those with non-diagnostic feature variables based on a single-mode distribution. We apply false discovery rate control to the feature selection procedure and provide a reasonable tradeoff between accuracy and efficiency. In the experimental study, the candidate features were generated based on the 3D shape derived from the landmarks on the specimen images from multiple views and local saliency characteristics from the 2D images directly. We evaluated the clustering accuracy and the relevance of the selected features using specimens in the genus *Carpiodes*. We found that two species, namely, *C. carpio* and *C. cyprinus*, are well separated from *C. velifer*, but do not form well separated clusters themselves. Interestingly, the previously misdiagnosed specimens from the Rio Grande and upper Colorado River in Texas are largely correctly grouped into the *C. cyprinus*-like cluster. The results show good potential for a computer-aided approach to taxonomic research and species diagnosis.

Appendix: Proof of FDR Control in Feature Selection

In the appendix, we show that the feature subset selection method presented in Section III controls the false discovery rate to be below the user specified level.

Proof: Denote by F the number of non-diagnostic features being selected and T the total number of features being selected. The FDR procedure intends to control the expected value of the random variable $Q = F/T$. Define $Q = 0$ if $T = 0$ since no error of false selection is committed. Let d_0 be the number of true non-diagnostic features. Denote by $p_1, ..., p_{d_0}$ the p-values corresponding to the non-diagnostic features. Without loss of generality, we assume that $1 \leq d_0 \leq d-1$. Let $d_1 = d - d_0$ and denote by $p^*_1, ..., p^*_{d_1}$ the p-values corresponding to the d_1 true diagnostic features. Denote the ordered p-values $p_{(1)} \leq ... \leq p_{(d_0)}$ and $p^*_{(1)} \leq ... \leq p^*_{(d_1)}$ corresponding to the non-diagnostic and diagnostic features, respectively. Define S to be the largest integer j satisfying $p^*_{(1)} \leq u_1, ..., p^*_{(j)} \leq u_j$. Let $S = 0$ when $p^*_{(1)} > u_1$. Clearly, S diagnostic features would be selected by the FDR controlled procedure if d_1 diagnostic features were given. Define the conditional error rate $q_e = E(Q|p^*_1, ..., p^*_{d_1})$. We will show next that $q_e \leq q$ for $1 \leq d_1 \leq d-1$ from

which the FDR is clearly below the significance level q.

$$\begin{aligned}
q_e &= E(F/T|p_1^*,...,p_{d_1}^*) \\
&\leq E[F/(S+F)|p_1^*,...,p_{d_1}^*] \ \text{(since } S+F \leq T\text{)} \\
&\leq \tfrac{d_0}{S+d_0} P(\min(p_1,...,p_{d_0}) \leq u_{S+1}) \\
&\quad \text{(since all nondiagnostic features are included)} \\
&\leq \tfrac{d_0}{S+d_0} \textstyle\sum_{i=1}^{d_0} P(p_i \leq u_{S+1}) \\
&\leq \tfrac{d_0}{S+d_0}(d-S)u_{S+1} \ \text{(since } d_0+S \leq d\text{)} \\
&= \tfrac{d_0(d-S)}{S+d_0} \min\left(1, \tfrac{d}{(d-S)^2} q\right) \\
&\leq \tfrac{d_0 d}{(S+d_0)(d-S)} q = \tfrac{d_0 d}{d_0 d + S(d-d_0-S)} q \\
&\leq q
\end{aligned}$$

ACKNOWLEDGMENT

We thank Prof. Yixin Chen at University of Mississippi for numerous discussions concerning this work.

REFERENCES

[1] A. Acero, J. J. Tavera, J. Reyes, "Systematics of the Genus Bagre (Siluriformes: Ariidae): A Morphometric Approach", *Cybium*, 29(2), pp. 127–133, 2005.

[2] D. C. Adams, F. J. Rohlf, D. E. Slice, "Geometric Morphometrics: Ten Years of Progress Following the 'Revolution' ", *Ital. J. Zool.*, 71, pp. 5–16, 2004.

[3] W. R. Atchley, B. K. Hall, "A Model for Development and Evolution of Complex Morphological Structures", *Biological Reviews* 66, pp. 101–157, 1991.

[4] R. C. Bailey, "Multivariate Analysis in Taxonomic Studies", *Canadian Journal of Fisheries and Aquatic Sciences*, 43, pp. 2054–2055, 1986.

[5] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society, Series B*, 57(1), pp. 289–300, 1995.

[6] R. Caruana and D. Freitag, "Greedy Attribute Selection", Proc. of *Int. Conf. on Machine Learning*, pp. 28-36, 1994.

[7] H. Chen, S. Huang, and H. L. Bart, "Taxonomy in Fish Species Complexes: A Role for Multimedia Information", *Int. Workshop on Multimedia Signal Processing*, Victoria, BC, Canada, Oct. 2006.

[8] Y. Chen, H. L. Bart, S. Huang, and H. Chen, "A Computational Framework for Taxonomic Research: Diagnosing Body Shape within Fish Species Complexes", Proc. of *Int. Conf. on Data Mining*, Houston, TX, Nov. 2005.

[9] S. Das, "Filters, Wrappers and A Boosting-Based Hybrid for Feature Selection", Proc. of *Int. Conf. on Machine Learning*, pp. 74-81, 2001.

[10] M. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature Selection for Clustering A Filter Solution", Proc. of *Int. Conf. on Data Mining*, pp. 115-122, 2002.

[11] S. Ferson, F. J. Rohlf, and R. K. Koehn, "Measuring Shape Variation of Two-Dimensional Outlines", *Systematic Zoology* 34(1), pp. 59–68, 1985.

[12] A. J. Haines, J. S. Crampton, "Improvements to the Method of Fourier Shape Analysis as Applied in Morphometric Studies", *Palaeontology*, 43(4), pp. 765–783, 2000.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2001.

[14] R. Kumar, P. Anandan, and K. Hanna, "Shape Recovery from Multiple Views: A Parallax Based Approach", *ARPA Image Understanding Workshop*, Monterey, CA, Nov. 1994.

[15] T Kadir, Scale Saliency and Scene Description, Ph.D. dissertation, University of Oxford, 2002.

[16] S. Lele, and J. T. Richtsmeier, "Euclidean Distance Matrix Analysis: A coordinate Free Approach for Comparing Biological Shapes Using Landmark Data", *American Journal of Physical Anthropology*, 86, pp. 415–428, 1991.

[17] T. Lindeberg, "Feature Detection with Automatic Scale Selection", *International Journal of Computer Vision*, 30(2), pp. 77–116, 1998.

[18] D. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge Press, 2003.

[19] J. Rissanen, "Stochastic Complexity and the MDL Principle", *Econometric Reviews*, 6, pp. 85–102, 1987.

[20] F. J. Rohlf, L. F. Marcus, "A Revolution in Morphometrics", *Trends in Ecology and Evolution*, 8, pp. 129–132, 1993.

[21] R. D. Suttkus, and H. L. Bart, "A Preliminary Analysis of the River Carpsucker, Carpiodes Carpio, in the Southern Portion of its Range," *L. Lozano (ed.) Libro jubilar en honor al Dr. Salvador Contreras Balderas*, Universidad Autonoma de Nuevo Leon, Monterrey Mexico, pp. 209–221, 2002.

[22] D. F. Wiley, N. Amenta, D. A. Alcantara, D. Ghosh, Y. J. Kil, E. Delson, W. Harcourt-Smith, F. J. Rohlf, K. St. John, B. Hamann, "Evolutionary Morphing", *Proceedings of IEEE Visualization*, 2005.

[23] M. Zelditch, D. Swiderski, D. Sheets, and W. Fink, *Geometric Morphometrics for Biologists: A Primer* Boston, Elsevier, 2004.

Huimin Chen received the B.E. and M.E. degrees from Department of Automation, Tsinghua University, Beijing, China, in 1996 and 1998, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, in 2002, all in electrical engineering. He was a post doctorate research associate at Physics and Astronomy Department, University of California, Los Angeles, and a visiting researcher with the Department of Electrical and Computer Engineering, Carnegie Mellon University from July 2002 where his research focus was on weak signal detection for single electron spin microscopy. He joined the Department of Electrical Engineering, University of New Orleans in Jan. 2003 as an assistant professor. His research interests are in general areas of signal processing, estimation theory, and information theory with applications to target detection and target tracking.

Henry L. Bart, Jr. is Professor of Ecology and Evolutionary Biology at Tulane University, and Director and Curator of Fishes of the Tulane Museum of Natural History. He is Editor of Tulane Studies in Zoology and Botany and Occasional Papers Tulane University Museum of Natural History. He earned BS and MS degrees from University of New Orleans and a Ph.D. (1985) in Zoology from the University of Oklahoma. He held faculty positions at the University of Illinois and Auburn University prior to joining Tulane University in 1992. His area of research specialization is ecology and systematics of freshwater fishes.

Shuqing Huang received the B.S. degree from Department of Computer Science, Zhongshan Univeristy, Guangzhou, China, in 1996, the M.S. degree from Department of Electrical Engineering and Computer Science, Tufts University, USA, in 2002, and the Ph.D. degree from Department of Electrical Engineering and Computer Science, Tulane University, USA, in 2007. She is currently a senior software engineer with General Dynamics Information Technology, New Orleans, USA. Her research interests include machine learning and data mining.