Taxonomy in Fish Species Complexes: A Role for Multimedia Information

Conference Paper · November 2006

DOI: 10.1109/MMSP.2006.285354 · Source: IEEE Xplore

3 authors, including:

Huimin Chen
University of New Orleans
98 PUBLICATIONS **1,851** CITATIONS
SEE PROFILE

Henry Bart
Tulane University
273 PUBLICATIONS **1,566** CITATIONS
SEE PROFILE

Taxonomy in Fish Species Complexes: A Role for Multimedia Information

Huimin Chen
Department of Electrical Engineering
University of New Orleans
New Orleans, LA 70148, USA
hchen2@uno.edu

Shuqing Huang
Department of Electrical Engineering & Computer Science
Tulane University
New Orleans, LA 70118, USA
shuang4@tulane.edu

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

***Abstract*—Biologists could make valuable use of the wealth of specimen information in natural history museum databases. "Taxonomy via the Internet" aims to build a centralized database where biologists can store, manipulate and retrieve biologically meaningful data from images of specimens and use the data to classify the specimens taxonomically. The major challenge in discovering and defining new species at present lies in (a) the scarcity of taxonomic expertise and (b) the tremendous effort involved in taxonomic research, as traditionally practiced. Multimedia information representation provides a new computational tool for extracting useful features from large databases of specimen images and has potential to expedite the pace of taxonomic research. In this paper, we use a taxonomic problem involving species of suckers in the genus Carpiodes to demonstrate the utility of this method. Logistic regression classifier with fully automated feature selection procedure is compared with the best landmark based classifier to illustrate how image quality affects classification accuracy. We discuss the need of creating a multimedia database using images from fish collection.**

***Keywords*—taxonomy; feature selection; logistic regression; shape analysis; multimedia representation**

***Topic area*—multimedia databases and application.**

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 the most update-to-date information on species identity, distribution and biology. Much of this information resides in natural history museums, inaccessible to most biologists. Most existing information systems of natural history are taxonomically focused (e.g., CephBase [1], FIGIS [2], FishBase [3], FishNet [4], HerpNet [5], MaNIS [6], Ornis [7]). They are designed to give the research community global access to specimen information within the taxonomic groups involved. However, they do not provide the most up-to-date information on species names, taxonomy and relationships.

The job of identifying and describing new species and determining interrelationships of species falls on taxonomists and systematists. Taxonomy and systematics, as traditionally practiced, can be painfully slow. One reason for this is that the number of taxonomic experts is small, especially for lesser known groups. Moreover, it takes tremendous amount of time and effort to examine and gather data from large numbers of specimens across broad geographical areas, particularly when working in developing countries. As a consequence, it is estimated that ninety percent of the world's species have yet to be discovered and described. The development of DNA sequencing technology and other molecular techniques has revolutionized systematics. However, the practice of taxonomy is still largely based on specimens.

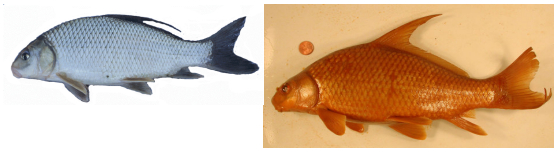
Multimedia information representation provides a new tool for dealing with the feature extraction from large numbers of specimens. It also creates new challenges in database usage and information retrieval, especially where taxonomists do not provide precise rules for multimedia-based recognition systems. In fact, most of the existing taxonomic databases contain only text and very few contain images directly digitized from the specimens. Taxonomists rarely rely on image processing and machine learning techniques to identify distinct features of species or for understanding relationships among species. On the other hand, computer scientists routinely deal with data mining and feature selection for classifying images from large databases. Thus, multimedia information may be particularly relevant to the process of taxonomic revision and new species discovery. In this paper, we use multimedia information derived from images of specimens to resolve a taxonomic problem in the fish genus *Carpiodes*, with and without expert knowledge of the body shape. More specifically, we compare the classification accuracy between a logistic regression

classifier with fully automated feature selection and the best landmark based classifier used in another study [14]. We illustrate how image quality affects the feature selection and classification accuracy and argue that the 2D body shape of a specimen may be inadequate to identify and formally describe the unrecognized species. Finally, we discuss the construction of 3D models to represent specimens and some of the consequences of using these models in database management.

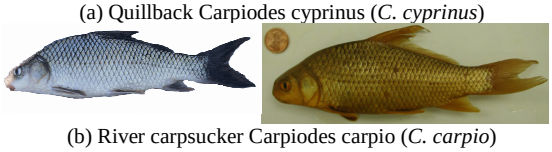
II. CARPIODES SPECIES COMPLEXES OVERVIEW

The genus of Carpiodes, as currently recognized, comprises three widely distributed species: the quillback *Carpiodes cyprinus*, the river carpsucker *Carpiodes carpio*, and the highfin carpsucker *Carpiodes velifer*[1]. Fig. 1 shows representative images of the three species and corresponding specimens. Taxonomists have long suspected that genus *Carpiodes* is more diverse than presently recognized and that each of the current species is a complex of multiple species in need of revision.

The classification of a specimen sample into one of the three species relies on the domain knowledge of the body features as summarized in Tab. 1 [12]. Computer technology has enabled the development of morphometric techniques to analyze the variation in body shape systematically using biologically definable measures, i.e., *homologous* landmarks, along the body outline [9]. Fig. 2 shows the positions of 15 landmarks digitized on a *Carpiodes* specimen using TpsDIG software [11]. Feature variables were derived from the 2D coordinates of these 15 landmarks using canonical variate analysis [12] and other feature selection methods. An important question is whether the three *Carpiodes* species complexes can be distinguished based on the body shape, i.e., a small set of selected features derived from the landmarks [14].

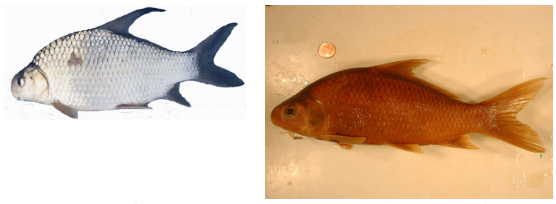


(a) Quillback Carpiodes cyprinus (*C. cyprinus*)



(b) River carpsucker Carpiodes carpio (*C. carpio*)

[1]http://www.funet.fi/pub/sci/bio/life/pisces/actinopterygii/cypriniformes/catostomoidea/catostomidae/ictiobinae/carpiodes/



(c) Highfin riversucker Carpiodes velifer (*C. velifer*)

Fig. 1. Three species of Carpiodes [12]

Table1. Distinct body characteristics of the three species

	head	body	snout	lip nipple
C. cyprinus	large	elongate	long	no
C. carpio	large	elongate	short	yes
C. velifer	small	short/deep	short	yes

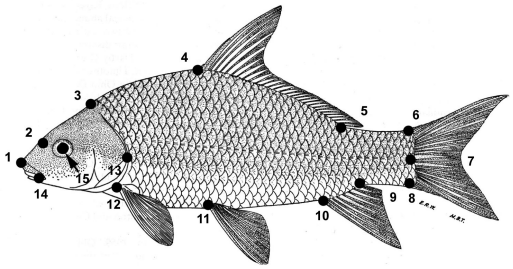


Fig. 2. A specimen with 15 landmarks made by domain experts [11]

III. JOINT FEATURE SELECTION AND CLASSIFICATION FOR CARPIODES SPECIES

A. The Taxonomic Problem

Shape analysis has been used to characterize variation in body proportions among *Carpiodes* populations across broad geographic areas [12]. However, the use of mophometric techniques alone can generate misleading results. Populations of *Carpiodes* in the Rio Grande and upper Colorado River in Texas have traditionally been identified as *C. carpio*. However, a recent DNA sequence analysis suggests that these populations have close affinities to *C. cyprinus* [12]. In [14], landmarked images of 650 *Carpiodes* specimens from Tulane Museum of Natural History (TUMNH) Fish Collection were used to identify features diagnostic for the 53 *Carpiodes* specimens from the Rio Grande and upper Colorado River. Over 50% of the specimens were correctly diagnosed as *C. cyprinus* based on two statistically significant feature variables, which is consistent with DNA sequence results [12]. However, several important issues still need to be addressed. First, putting 15 landmarks on 2D images of 650 specimens is a laborious task. A complete analysis of *Carpiodes* specimens in the TUMNH Fish Collection alone would involve digitizing landmarks on over 20,000 *Carpiodes*

specimens. Second, the features used to diagnose *Carpiodes* specimens from the Rio Grande and upper Colorado River are different from those used in visual determination by a domain expert. The main character used in [12] to differentiate Rio Grande and upper Colorado River *Carpiodes* specimens from *C. carpio* was the absence of a lower lip nipple, which could not be derived from landmark data. One may question whether a 2D digital image of a specimen is adequate to diagnose the body shape differences when fish are 3D in actuality. Third, what does it mean biologically if the classifier fails to place most of the undetermined specimens into one of the known species? Is this a strong indication that the specimens in a group need to be considered as a new species?

B. Feature Selection with Controlled False Discovery Rate

To address the first question, we compare feature selection with and without landmark information. Without loss of generality, we denote $\mathbf{x}_i$ as feature vector of the *i*-th specimen which can be obtained either through the 2D coordinates of the landmarks or directly from the 2D image. The number of features is usually large and may not all be useful for classification purposes. In fact, some features are highly correlated and the experimental study in [14] showed that as few as two good features can yield fairly accurate classification results for a 1-norm support vector classifier. Feature selection can also be viewed as a multiple hypothesis testing problem. Assume the number of features is *d* and a hypothesis H_k corresponds to the selected index set $I_k \subseteq \{1,...,d\}$. One needs to find the best hypothesis to separate the known species. The number of hypothesis grows exponentially in *d*. One viable solution is to select the subset of features with a controlled false discovery rate (FDR) [13]. The procedure requires to have test statistics T_1, …, T_d for all features with the associated *p*-values indicating the statistical significance. Then for any user specified FDR $q \in (0,1)$, the feature selection is carried out by the following steps [13]:

- Order the *p*-values such that $p_{(1)} \le ... \le p_{(d)}$.
- Compute the index $k = \max\{i \mid p_{(i)} \le iq/d\}$.
- Select *k* features corresponding to the ordered *p*-values $p_{(1)},..., p_{(k)}$.

The scheme is very efficient and has many good theoretical properties when *q* is chosen to be very small [8].

The difficulty of obtaining an accurate *p*-value for each feature lies in that the sample size of the known species is fairly small compared with the number of candidate features. We use logistic regression to tackle the feature selection and classification problem jointly. Details can be found in [14].

C. Logistic Regression Classifier

Logistic regression classifier (LRC) is a computationally efficient algorithm to handle a large number of feature variables [16]. For simplicity, we consider a binary classification problem with the following assumed statistical model.

$$P(y_i = 1 \mid \mathbf{x}_i, \mathbf{w}) = 1/(1+e^{-\mathbf{w}^T \mathbf{x}_i})$$

where $\mathbf{w}$ is the regression coefficient vector to be estimated from the training samples. The log-likelihood of *N* independent training samples is given by $L(\mathbf{w}) = \sum_{i=1}^{N} \left\{ y_i \mathbf{w}^T \mathbf{x}_i - \log(1+e^{\mathbf{w}^T \mathbf{x}_i}) \right\}$. An efficient algorithm to obtain the maximum likelihood estimate of $\mathbf{w}$ is through iteratively re-weighted least squares method [16]. The *p*-value for each component of $\mathbf{w}$ can also be obtained conditioned on the hypothesis that it is zero. Thus feature selection can be directly applied to the above logistic regression classifier.

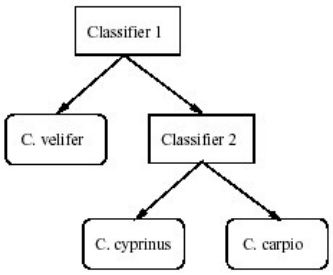


Fig. 3. Best classification tree for genus *Carpiodes* [14]

For the multiple class case, we build a classification tree using logistic regression as the binary classifier and select the tree with the best classification accuracy using all training samples. With landmark information, the best hierarchical tree structure is shown in Fig. 3 where the 53 undetermined specimens are not used [14]. The leading feature used by the logistic regression classifier is the distance between the naris and the tip of the snout in proportion to the distance between the naris and the eye which is related to the size and shape of the head. The best feature for diagnosing *C. cyprinus* from *C. carpio* is the slope of the line connecting the naris and the back of the mouth, which is related to the size of the head relative to the size and position of the mouth. The two species are otherwise similar in overall

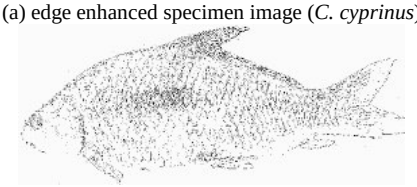
body shape. Another feature that unites *C. carpio* with *C. velifer*, but distinguishes it from *C. cyprinus* is the presence of a lip nipple. However, this character can not be seen in 2D images based on the side views of the specimens. Thus, classification accuracy of specimens of *C. carpio* and *C. cyprinus* is worse than that of the first level of the classification tree.

D. Feature Selection without Using Landmarks

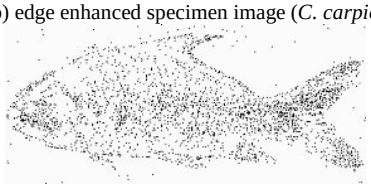
Image based object recognition is a challenging research problem in its own right. Here we try to find useful features from the edge enhanced and normalized grey image based on the detected saliency regions [18]. A saliency region indicates high complexity of signal intensity measured by the entropy. The saliency region detector works as follows. For each pixel location, a probability density function (PDF) of the signal in a circular region of radius (scale) *s* is estimated. The signal is the intensity value of the equalized grey scale image. The PDF is approximated by a signal histogram computed over a circular region. Then the entropy, *H(s)*, is calculated for each scale. Those scales at which *H(s)* is a maximum are chosen to be candidate scales. The saliency of each candidate is evaluated using a measure of the self-dissimilarity in the scale-space [21]. Regions with saliency value greater than a threshold are selected as the salient regions. Each salient region is defined by its center and the radius. The salient regions are used as the candidate feature variables for the logistic regression classifier with an FDR controlled feature selection procedure. Since it is computationally expensive to find the saliency regions for a high resolution image, we also consider an edge detection based method to lower the image quality with coarse quantization on the level of grey scales. We used a Sobel operator [15] with an increasing threshold τ to keep only the strong edges of the image and then applied the saliency region detector to the compressed image. An example of the compressed low quality images of three specimens is shown in Fig. 4 with τ=0.02.

We compare the classification accuracy with the best landmark based classifier presented in [14] using the training samples of 297 *C. cyprinus* specimens, 128 *C. carpio* specimens, 172 *C. velifer* specimens and the test samples of 53 specimens from the Rio Grande and upper Colorado River. The results for separating *C. velifer* from *C, carpio* and *C. cyprinus* are listed in Tab. 2. We can see that the saliency based classifier achieves comparable accuracy to the landmark based classifier with no undetermined specimens being misdiagnosed as *C. velifer*. As the threshold increases, the image quality declines and consequently the classification error increases. This is also evidenced from Fig. 4 where the classification by visual examination from a domain expert is challenging without knowing the species of each image.

(a) edge enhanced specimen image (*C. cyprinus*)



(b) edge enhanced specimen image (*C. carpio*)



(c) edge enhanced specimen image (*C. velifer*)

Fig. 4. Three Specimens of *Carpiodes* with Edge Enhanced Image Compression

Table 2. Separating velifer from cyprinus and carpio: landmark based classifier vs. saliency region based classifier

algorithm	training error	testing error
LRC/landmark	5.5%	0%
LRC/τ=0.001	5.7%	0%
LRC/τ=0.005	6.7%	1.9%
LRC/τ=0.02	10.7%	9.4%

Next, we compare the classification accuracy for separating *C. cyprinus* from *C. carpio*. The results are listed in Tab. 3. Again, the saliency based classifier achieves comparable accuracy in training and both classifiers put most of the undetermined specimens into *C. cyprinus*, which is a strong indication that the undetermined specimens do not belong to *C. carpio*. Interestingly, as the image quality declines, more specimens are classified as *C. carpio*. This is mainly because the edge enhancement preserves the overall shape information which alone diagnoses the specimens as *C. carpio* incorrectly (see [14] for details). The results indicate that the FDR controlled feature selection for a logistic regression classifier performs equally well even without landmark information although the computational load is much more expensive without doing image compression.

Table 3. Separating *C. cyprinus* from *C. carpio*: landmark based classifier vs. saliency region based classifier

algorithm	training error	testing
LRC/landmark	8.0%	22.6%
LRC/τ=0.001	5.2%	39.6%
LRC/τ=0.005	13.2%	45.3%
LRC/τ=0.02	17.4%	54.7%

E. Statistical Significance Test for New Species Diagnosis

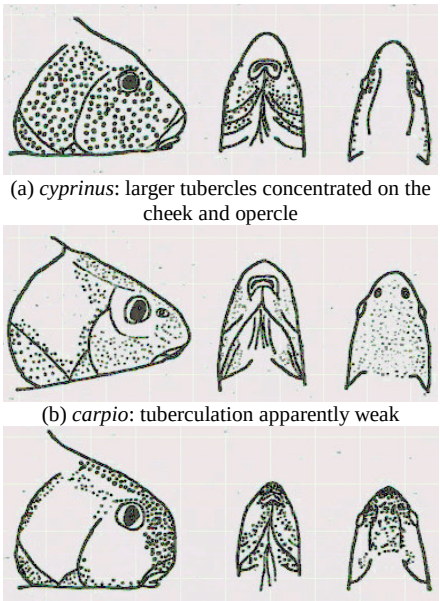
In diagnosing the 53 undetermined specimens, the classifiers using landmarks or saliency regions can not put the majority of the specimens into one of the known species, which is a strong indication for the domain expert to perform further diagnosis in order to resolve the taxonomy. We randomly pick 53 specimens from *C. cyprinus* and use the remaining samples to train the classifiers. The significance test results are listed in Tab. 4 where we can see that both classifiers with and without landmark information yield reasonably large *p*-values so that the hypothesis that the testing specimens belong to cyprinus should not be rejected. Note that as the image quality reduces, the classification error increases and one observes small *p*-values which may generate false diagnosis. Thus caution has to be exercised when choosing a threshold of the *p*-value especially for large training error.

Table 4. Significance test: landmark based classifier vs. saliency region based classifier

algorithm	training error	testing (*p*-value)
LRC/landmark	6.2%	0.13
LRC/τ=0.001	5.7%	0.17
LRC/τ=0.005	9.5%	0.02
LRC/τ=0.02	16.9%	0.008

IV. TOWARD 3D SHAPE ANALYSIS

Currently recognized species of *Carpiodes* are mainly based on body shape characteristics and it seems that machine learning techniques can provide informative diagnosis with 2D images of the specimens. However, the selected features are difficult to measure and hard to interpret, which makes classification results questionable. In fact, some distinct features can not be identified from the 2D image of the specimen but are useful for diagnostic purposes. For example, it has long been known that breeding tuberculation patterns can be used to diagnose *Carpiodes* species (see Fig. 5 [17]). However, the pattern must be assessed from dorsal and ventral, as well as lateral views of specimens. These views are unavailable in a lateral 2D image. The same is true on the lip nipple character (visible only in ventral view) that distinguishes *C. carpio* and *C. velifer* from *C. cyprinus*. For these characters and to properly represent the depth dimension, it would be more accurate and appropriate to have multiple views of each specimen or to create 3D visualization based on these images, and make them accessible through the Internet. One possible way to visualize 3D shape is to use skeletons as a descriptor of the shape. The general idea is to derive 1D skeletal curves from a 3D object such that each curve represents a significant part of the object. These curves are then converted into an attributed graph representation called a skeletal graph [22]. It can then be used for indexing, matching, correspondence finding and other semantic queries.



(a) *cyprinus*: larger tubercles concentrated on the cheek and opercle

(b) *carpio*: tuberculation apparently weak

(c) *velifer*: tubercles concentrated on the snout and top of head

Fig. 5. Breeding tuberculation pattern of *Carpiodes* [17]

V. SHAPE-BASED INFORMATION RETRIEVAL

As we have seen, the 2D image-based specimen representation may not be adequate for species diagnosis especially when the views of important diagnostic feature are missing. The problem with using 3D models for each specimen is that the database management and information retrieval techniques need to be developed in accordance with these models. The key issue in developing a shape-based information retrieval and analysis system is to find a shape descriptor for which an index can be built; similarity queries can be answered efficiently so that feature selection and classification can be implemented similar to the 2D case. One possible solution is to construct a

shape distribution sampled from a particular function measuring the geometric properties of a 3D model. A set of functions have been proposed such as reflective symmetry descriptor [19] and spherical harmonics [20]. Features derived from the shape descriptor can have better discriminative capability for species diagnosis. One important issue for building taxonomic trees based on 3D specimen samples is how to organize the database to support shape-based user queries. The simplest query interface is to search for 3D objects based on textual keywords. A more advanced requirement is to support shape-based queries such that shape similarity can be used for taxonomic purposes, enabling taxonomists to easily obtain the 3D information of a specimen online. Such a search engine has been developed in Princeton Shape Benchmark [24] with a significant number of users.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we compared the classification accuracy for a logistic regression classifier using false discovery rate controlled feature selection with and without landmark information using a taxonomic problem in the genus *Carpiodes*. We found that classification without landmark information, albeit computationally more expensive, has comparable performance to that using landmarks. The classification accuracy degrades as we reduce the image quality by preserving only strong edges, which indicates the importance of local shapes for diagnostic purposes. To properly visualize specimens and a more complete set of diagnostic features, we propose the use of 3D shape models to characterize each specimen and discuss its impact on the feature selection and classification algorithm used for the 2D images. Finally, we briefly highlighted the representation of 3D objects and the impact on database management and information retrieval.

There are many avenues for future work. The logical next step is to build 3D models of specimens and test the accuracy of species diagnosis. The results should be compared with those obtained using 2D images to identify the important views of a 3D specimen for taxonomic problems. Another important direction is to design a database system for the 3D images that allows users to search and mine specimens via the Internet. Efforts are underway to build virtual natural history cyber-laboratories, thereby accelerating the rate of new species discovery [23].

REFERENCES

[1] CephBase, http://www.cephdev.utmb.edu/.
[2] FIGIS, Fisheries Global Information System, http://www.fao.org/figis/servlet/static?dom=root&xml=index.xml.
[3] FishBase, http://www.fishbase.org/search.php.
[4] FishNet, http://speciesanalyst.net/fishnet/default.html.
[5] HerpNet, http://www.herpnet.org/.
[6] Mammal Networked Information System, http://elib.cs.berkeley.edu/manis/.
[7] Ornithological Information System, http://ornisnet.org/
[8] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone, "Adapting to Unknown Sparsity by Controlling the False Discovery Rate", *Annals of Statistics*, 2005.
[9] D. C. Adams, F. J. Rohlf, D. E. Slice, "Geometric Morphometrics: Ten Years of Progress Following the 'Revolution' ", *Ital. J. Zool.*, 71:5-16, 2004.
[10] N. Amenta, S. Choi, and R. Kolluri. "The Power Crust", in the 6th *ACM Symposium on Solid Modeling and Applications*, 2001.
[11]Geometric Morphometrics of *Carpiodes* Species Complexes, Systematics of Ictiobinae, 2002, http://www.museum.tulane.edu/ictiobin/morphometrics.html
[12] H. L. Bart, M. D. Clements, R.E. Blanton, M. Cashner, K. R. Piller, and D.L. Hurley, "Incongruence of mtDNA Sequence and Morphological Evidence in *Carpiodes* Species Complexes (Teleostomi: Catostomidae)", 84th Annual Meeting, *American Society of Ichthyologists and Herpetologists*, Norman, Oklahoma, 2004.
[13] 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):289-300, 1995.
[14] Y. Chen, H. L. Bart, S. Huang, and H. Chen, "A Computational Framework for Taxonomic Research: Diagnosing Body Shape within Fish Species Complexes", submitted to *Int. Conf. on Data Mining*, 2005.
[15] R. C. Gonzalez, and R. E. Woods, *Digital Image Processing*, 2nd Ed., Addison-Wesley, 1992.
[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2002.
[17] G. R. Huntsman, "Nutial Tubercles in Carpsuckers (Carpiodes)," *Copeia*, 1967(2):457-458, 1967.
[18] T. Kadir and M. Brady, "Scale, Saliency and Image Description," *International Journal of Computer Vision*, 45(2):83-105, 2001.
[19] M. Kazhdan, B. Chazelle, D. Dobkin, T. Funkhouser, and S. Rusinkiewicz, "A Reflective Symmetry Descriptor for 3D Models", *Algorithmica*, 38(2):201-225, November 2003.
[20] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors", *Symposium on Geometry Processing*, 2003.
[21] T. Lindeberg, "Feature Detection with Automatic Scale Selection", *International Journal of Computer Vision*, 30(2):77-116, 1998.
[22] G. Malandain, and S. Fernandez-Vidal, "Euclidean Skeletons", *Image and Vision Computing*, 16:317-327, *1998*.
[23] L. Page, H. Bart, R. Beeman, L. Bohs, L. Deck, V. Funk, D. Lipscomb, M. Mares, L. Prather, J. Stevenson, Q. Wheeler, J. Wooley, D. Stevenson, *LINNE: Legacy Infrastructure Network for Natural Environments*. Illinois Natural History Survey, Champaign, 2005.
[24] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The Princeton Shape Benchmark", *Shape Modeling International*, Genova, Italy, June 2004.