Fine-Grained Recognition for Arthropod Field Surveys: Three Image Collections

Jeffrey Lin[1], Natalia Larios[2], David Lytle[1], Andrew Moldenke[1], Robert Paasch[1], Linda Shapiro[2], Sinisa Todorovic[1] and Tom Dietterich[1]

[1]Oregon State University
Corvallis, Oregon, USA 97331
tgd@cs.orst.edu

[2]University of Washington
Seattle, Washington, USA 98195
http://web.engr.oregonstate.edu/~tgd/bugid/

Abstract

Many research groups and governmental organizations, including the US Environmental Protection Agency, collect samples of insect larvae from freshwater streams to assess the health of stream ecosystems. Specimens in the samples are manually identified to the level of species or species group and counted. This is expensive, time-consuming, and requires many years of experience. Automating this visual identification task requires highly-accurate fine-grained recognition methods. This abstract describes three databases of high-resolution images created to promote the development of such methods, presents benchmark results on these images, and discusses some of the issues raised for fine-grained recognition.

1. Introduction

Aquatic biomonitoring is a methodology for assessing the health of freshwater ecosystems by examining the organisms that live there. It is an important tool for basic research, pollution monitoring, and evaluating restoration efforts. One commonly-used group of organisms for biomonitoring are the EPTs: Ephemeroptera (Mayflies), Plecoptera (Stoneflies), and Trichoptera (Caddis flies). The EPTs, often referred to as macroinvertebrates, range in size from 0.1-10mm.

Little training is required to collect and clean EPT samples, but years of experience are needed to learn to identify specimens to species level. When identifying a specimen, experts typically manipulate it to examine the mouth and genitalia, since these provide valuable discriminating features. However, our collaborators report that with enough experience, a single glance at the specimen is enough to give them a pretty good idea of the species. Hence, it is reasonable to expect that computer vision methods can automate much of the identification task.

Over the past 8 years, we have collected three sets of samples, manually identified each specimen, and photographed the specimens to produce three image databases: STONEFLY9 [4], EPT29, and EPT54. STONEFLY9 consists of 3826 images 773 specimens from 9 taxa (species or genera) of stoneflies. EPT54 consists of 10,173 images of 3394 specimens belonging to 54 taxa of EPTs; EPT29 is a subset of 4722 images of 1596 specimens of 29 taxa from EPT54. Each specimen was photographed multiple times using a semi-automated apparatus under fixed lighting, focus, and exposure conditions. The images are captured at high resolution (2560 x 1920 pixels). The version of the apparatus employed for EPT29 and EPT54 automatically positions each specimen on its back and then photographs it from below. Each object is segmented from the (standard) background via Bayesian matting and morphological operations. Then PCA is applied to identify the principal axis, and an SVM classifier is applied to orient the specimen so that it is facing right. Figure 1 shows examples of the specimens.

2. Classification Task

Although each specimen is photographed multiple times, a reasonable research task is to treat each image as a separate data point for classification. To ensure unbiased evaluation, all images of a single specimen must be kept together in the same training set, validation set, or test set. In our experiments, we employed a 3-fold cross-validation, which allowed one fold for training, one for auxiliary tasks (e.g., dictionary construction, parameter tuning), and one for testing. Relevant performance metrics are overall error rate and rejection rate at 90%, 95%, and 99% precision. In an application setting, specimens rejected by the classifier must be manually identified, so the goal is to achieve a desired level of precision while minimizing the amount of manual work.

For each database, various descriptors have been extracted and can be downloaded in addition to the images themselves: STONEFLY9: SIFT at interest points and regularly sampled; EPT29 and EPT54: Beam Angle Descriptor computed at salient points on the perimeter of each specimen, HOG regularly sampled, SIFT at DoG interest points.

3. Classifiers

Three classifiers have been developed for this task Boosted dictionaries [1], Stacked evidence trees [2], and Stacked spatial pyramid SVM [3] and compared to stand-

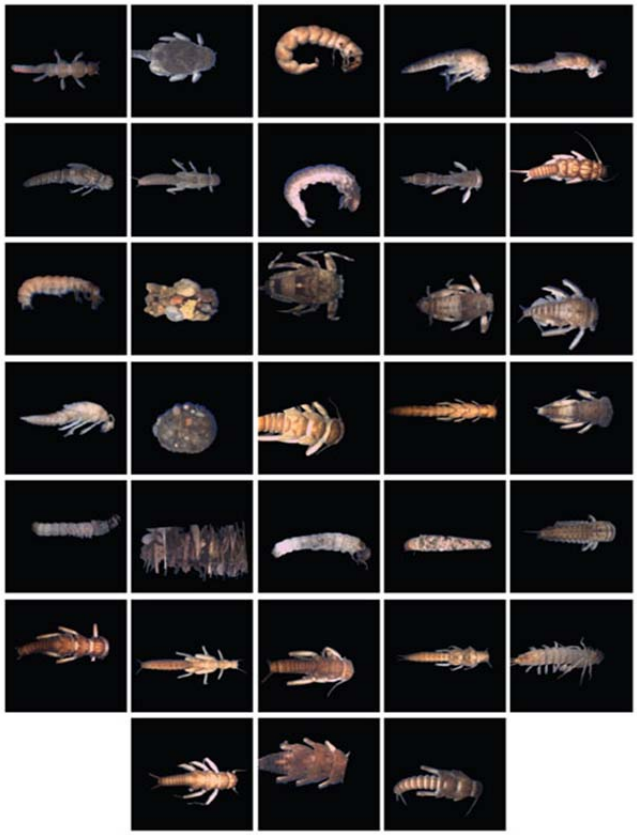

Figure 1: Oriented EPT29 Specimens. Caddisflies build and live in cases, so for some species this figure shows both the insect and the case (column 2, rows 3-5).

ard classifiers.

In the boosted dictionary method, each image is assigned an initial weight of 1, weighted K-means clustering is applied to construct a dictionary, each image is re-represented as a keyword histogram, and a classifier is trained. Then the images are reweighted according to the standard Adaboost method, and the process of dictionary construction and classifier training is repeated.

The Stacked Evidence Tree classifier trains evidence trees (a kind of random forest) to classify each descriptor vector separately. The results of those classification decisions (summarized as counts by class of the number of training examples reaching each leaf) are provided as input to a stacked classifier, which makes the final decision.

Similarly, the Stacked Spatial Pyramid classifier first applies random forests to classify each descriptor separately. The predicted class probabilities from these individual descriptors are aggregated according to a 3-level (16-4-1) spatial pyramid and provided as input to an SVM classifier using the spatial pyramid kernel, which makes the final decision.

Note that all three of these methods attempt to overcome the problem of information loss that occurs when standard dictionary methods are applied. By boosting dictionary construction, the first algorithm allocates more dictionary "resolution" where it is needed. The other two methods do not employ dictionaries, but instead operate directly on the descriptors.

4. Results

On STONEFLY9, the Boosted Dictionary classifier achieves 95.1% correct classification, Stacked Evidence Trees achieve 94.4%, and a baseline Gaussian mixture model dictionary combined with boosted decision trees gives 83.9% accuracy.

On EPT29, the Stacked Spatial Pyramid achieves 88.06% correct. The best conventional approach was an SVM using a global HOG with the χ^2 kernel, which achieves 68.21% accuracy.

On EPT54, the Stacked Evidence Trees achieve 74.3% accuracy. The rejection rate at 90% precision is 0.38 and at 95% precision is 0.58. Figure 2 shows the precision as a function of the fraction rejected.

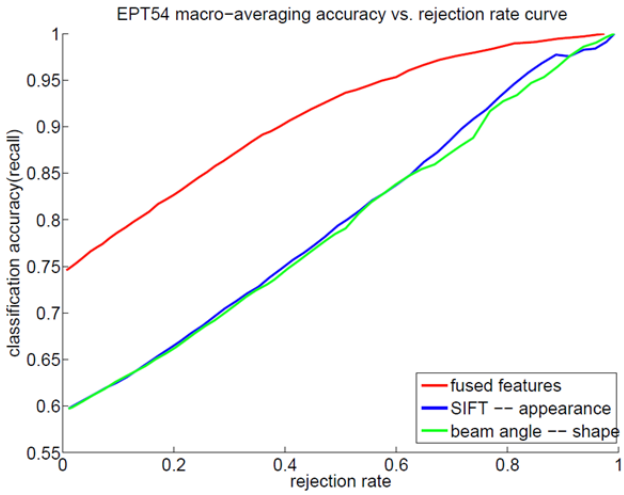

Figure 2: Precision versus rejection rate.

Acknowledgement

The authors gratefully acknowledge the support of the US NSF under Grants 0326052 and 0705765.

References

[1] Zhang W, Surve A, Fern X, Dietterich T. Learning Non-Redundant Codebooks for Classifying Complex Objects. In: *ICML-2009*.; 2009:1241-1248.

[2] Martinez G, Zhang W, Payet N, et al. Dictionary-Free Categorization of Very Similar Objects via Stacked Evidence Trees. In: *CVPR-2009*. IEEE; 2009:1-8.

[3] Larios N, Lin J, Zhang M, et al. Stacked Spatial-Pyramid Kernel: An Object-Class Recognition Method to Combine Scores from Random Trees. In: *IEEE WACV*; 2011.

[4] Dietterich, T. The STONEFLY9 Image Dataset. http://web.engr.oregonstate.edu/~tgd/bugid/stonefly9/