Stacked Spatial-Pyramid Kernel: An Object-Class Recognition Method to Combine Scores from Random Trees

N. Larios[*], J. Lin[†], M. Zhang[‡], D. Lytle[†], A. Moldenke[†], L. Shapiro[*], T. Dietterich[†]
[*] University of Washington, [†] Oregon State University, [‡] University of California, San Diego
[*] {nlarios@u, shapiro@cs}.washington.edu, [‡] mezhang@ucsd.edu
[†] {linju@eecs, lytleda@science, moldenka@science, tgd@eecs }.oregonstate.edu

Abstract

The combination of local features, complementary feature types, and relative position information has been successfully applied to many object-class recognition tasks. Stacking is a common classification approach that combines the results from multiple classifiers, having the added benefit of allowing each classifier to handle a different feature space. However, the standard stacking method by its own nature discards any spatial information contained in the features, because only the combination of raw classification scores are input to the final classifier. The object-class recognition method proposed in this paper combines different feature types in a new stacking framework that efficiently quantizes input data and boosts classification accuracy, while allowing the use of spatial information. This classification method is applied to the task of automated insect-species identification for biomonitoring purposes.

The test data set for this work contains 4722 images with 29 insect species, belonging to the three most common orders used to measure stream water quality, several of which are closely related and very difficult to distinguish. The specimens are in different 3D positions, different orientations, and different developmental and degradation stages with wide intra-class variation. On this very challenging data set, our new algorithm outperforms other classifiers, showing the benefits of using spatial information in the stacking framework with multiple dissimilar feature types.

1. Introduction

Object-class recognition is one of the main research areas in computer vision. Its goal is to understand and implement, in machine vision systems, the human ability of recognizing the abstract class to which a previously unseen object belongs. We extend the state-of-the-art recognition approaches that use a set of random trees to discriminatively structure the information provided by the feature data obtained from the training and testing images. The random trees stage replaces unsupervised cluster model learning and cluster assignment used in visual dictionary methods. The use of decision tree ensemble methods [9, 12, 15] has replaced the standard approach of employing unsupervised visual dictionaries [8] as the initial stage to structure and/or quantize the input features in classification methods. An intermediate approach is the creation of quasi-supervised [14] or unsupervised dictionaries that have some keywords eliminated later by discriminative measures.

In the work presented in this paper, we pursue an approach that retains the simplicity and elegance of methods such as the evidence tree [12], while employing the local-feature spatial information in a robust way. The use of spatial information has proven useful in visual dictionary [10] and tree ensemble methods [9, 15] for generic object recognition and image classification tasks. Our work uses a spatial pyramid-kernel SVM classifier [10] while allowing diverse feature types that can complement each other to be combined, and it has been shown to be a successful technique to boost classification accuracy [1, 12]. Our classification framework is able to consider the discriminative structures from different objects parts invariant to changes in positions and size. We illustrate our object-class recognition by focusing on a relevant environmental protection application, implementing automated insect-species identification in order to obtain biodiversity measurements that support and enable biomonitoring.

Insect-Species Classification Biomonitoring is the assessment of the status and trend of the environment using counts of a set of defined species known for their susceptibility and capacity to accumulate the effects of environmental changes over time. One of the most widely used biological monitoring metrics for water quality assessment is the population count of specimens from the insect orders: Ephemeroptera (mayflies), Plecoptera (stoneflies), and Trichoptera (caddisflies), often abbreviated as EPT. Species within these orders vary greatly in their susceptibility to pollution and so are robust indicators of stream water quality. Identifying and enumerating EPT samples is often a labor-intensive and time-consuming process, generally involving expert entomologists performing manual classification. This paper describes our method applied to insect-species identification and its application and evaluation on an image data set collected to emulate the species distribution found in typical EPT biomonitoring samples. This set of EPT specimen images contains individuals from 29 species, and thus it is referred as the *EPT29* dataset.

Related Work Previous work in automated species recognition for biological monitoring of water quality have only included subsets of the species commonly utilized in EPT measuring indexes. In [9, 12] experiments on nine stonefly taxa (Plecoptera) obtained very low classification errors. Earlier research [8] from the same group of authors evaluated the feasibility of automated stonefly species identification with existing computer vision and machine learning methods. Other work on biological water quality assessment has been focused on algae specimens [16]. In this series of studies regarding water quality assessment, the progression of recognition methods from simple pattern-recognition approaches to unsupervised and later discriminative dictionaries can be seen. Orchard pest control [17] is another application of automated insect identification. In this case, local and global features were combined to identify pest species. In [2, 7, 13] generic object recognition methods were applied to the recognition of winged insects. As part of the trend to develop mobile applications, an automated system for on-field identification of botanical species [18] using leaf shape and venation patterns was implemented.

Contribution Our method enables the combination of dissimilar local features generated by different detectors and descriptors in a discriminative framework. A diverse set of features is an important factor in our current application given wide inter-species appearance variation due to different 3D specimen position and orientation, natural diversity, developmental stages, and degradation after capture. We use shape and edge keypoints and a dense grid of patches as means to obtain useful regions for local features, which are described by the HOG and SIFT appearance descriptors and by the beam-angle histogram for shape. The relative position information represented by a spatial histogram is generated by the initial random trees stage using the appearance and shape feature scores. This combination of features and the direct use of spatial information are proved successful by a large classification accuracy increase in the challenging *EPT29* identification task.

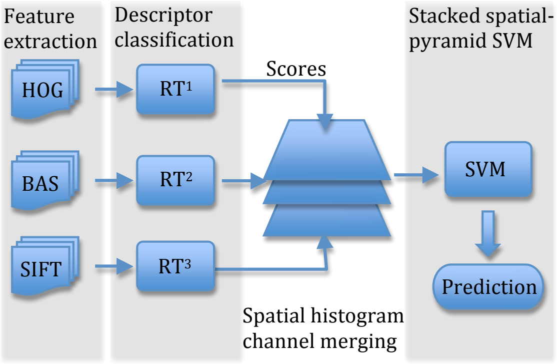
Feature extraction
Descriptor classification
Stacked spatial-pyramid SVM
HOG
BAS
SIFT
RT¹
RT²
RT³
Scores
SVM
Prediction
Spatial histogram channel merging

Figure 1. Overview of the classification architecture.

2. Stacked Spatial Classifier for Object-Class Recognition

The proposed classification architecture draws on the idea of using evidence trees for dictionary-free classification [12] and on the use of spatial information at the final classification level [9] to identify the species of the specimen. Insect identification research in [8, 12] has shown that combining multiple interest region detectors and their respective descriptors gives better results than single detectors. The benefit of feature combination has also been shown to work well in stacking methods for object recognition [1]. The need for a diverse set of features becomes even more evident when we consider the large variation present in all 29 species chosen for our current dataset. It becomes obvious that a combination of features is necessary, because no single feature type would distinguish between all of them. These characteristics make this recognition problem very challenging.

2.1. Overall Class Prediction

We now overview the prediction process of our method. For the *EPT29* dataset the image background is first segmented out and the specimen is automatically aligned with the horizontal axis of the images. The specimens then are oriented facing left by a linear-kernel SVM classifier evaluated on a global HOG [6] descriptor trained to distinguish between specimens of this dataset facing left or right. Samples of the resulting images are shown in rows (a) and (c) of Figure 3. After this preprocessing step, our classification framework is composed of three main stages: (1) region detection and descriptor extraction of low-level local features, (2) local-feature classification and spatial histogram computation, and (3) insect-species prediction. The overall prediction framework of our system is shown in Figure 1.

The initial classification stage of the system is composed by a set of C classifiers that match every (detector, descriptor) feature type employed. $\mathcal{Q}$ is the set of (detector, descriptor) pairs that are applied to each new image I to be evaluated. The stacking component of our method enables the combination of very different types of pairings, each with different invariant characteristics as well as the retention of their spatial information. In this work the employed pairs are: (1) salient points of high curvature [5] with a beam angle descriptor [3], (2) dense grid of overlapping image patches with the HOG descriptor, and (3) the difference of Gaussians detector and SIFT descriptor [11]. The application of each pair in $\mathcal{Q}$ generates a set of detections represented by a set of descriptor vectors $B^c_I = \{x^c_{I,1}, \ldots, x^c_{I,N_c}\}$ where N_c is the number of detections obtained by the (detector, descriptor) pair c.

The set of initial local-feature classifiers is composed of a random-trees classifier RT^c for each combination c. These classifiers are employed to obtain a probability classification score of each of the M classes for every descriptor vector $x^c_{I,j}$. The probability score $p \in \mathbb{R}^M$ is a M-dimensional vector employed to build the spatial histogram of classification scores $H^{c,L}_I$ of the finest spatial-grid resolution level L. This histogram is then used to construct a sequence of histograms with grid density-levels $\ell = 0, \ldots, L$ in the same manner as in [10]. Each histogram $H^{c,\ell}$ with resolution level ℓ has 2^ℓ cells along every spatial dimension (a total of $2^{2\ell}$ cells) and M channels in every cell $H^{c,\ell}(i)$. The set of histograms $\{H^{c,\ell}_I | c = 1, \ldots, C\}$ obtained with the set of random trees $\{RT^c\}$ are merged along the class-channel dimension into a singe spatial histogram H^ℓ_I of $C \times M$ channels and $2^{2\ell}$ spatial cells. The sequence of histograms $\{H^\ell_I | \ell = 0, \ldots, L\}$ constitutes the feature vector of image I for the final classifier, which is well suited for the spatial-pyramid kernel.

2.2. Combining Local-Classification Scores

The procedure to generate the histogram $H^{c,L}_I$ of image I is as follows. For every classifier c of $\mathcal{Q}$, every descriptor $x^c_{I,j}$ of B^c_I is evaluated by each tree in RT^c. Every tree votes for the class of the descriptor. The class probability score $p^c_{I,j}$ is computed by summing the votes for each class m in the vector component $p^c_{I,j}[m]$. The values of $p^c_{I,j}$ are then normalized to sum to 1, approximating the posterior probability $\mathbf{p}(m|x^c_{I,j})$. After the scores are obtained for every element in B^c_I, for each spatial cell i at grid density L, all the score vectors $\{p^c_{I,j} | l^c_{I,j} \in i\}$ whose descriptor location $l^c_{I,j}$ falls within the ith cell grid are accumulated with vector addition in the respective channels of bin $H^{c,L}(i)$. The accumulation of discriminative information in the form of scores is the main difference between our method and the original spatial-pyramid kernel SVM method [10], which only accumulates cluster assignment counts. These histograms carry the spatial information of local features to the stacked classifiers, which contrasts with [1], where spatial information is only employed at the first classification level. Thus our method has all the benefits of the bag-of-features approach while accumulating descriptor classification scores.

The use of the spatial histograms also has the advantage of indirectly maintaining information about the number of features detected in different image regions, which correlates with the image content. The results presented in Section 4.3 using the histograms generated with a single pair c compared with the ones obtained by combining all pairs in $\mathcal{Q}$ show that as with standard stacking techniques, our method benefits from combining complementary feature types in H^L_I.

2.3. Stacked Spatial-Pyramid Kernel SVM

In order to perform the pyramid matching in two-dimmensional image space, a sequence of histograms $\{H^\ell_I\}$ matching the grids at different resolution levels $\ell = 0, \ldots, L$ is built. Let $\jmath_{1,\ldots,4}$ be the grid cells at level $\ell + 1$ that subdivide cell i at level ℓ. The recursive process to compute the sequence of histograms starts with H^L_I. All the bins in cell i at level ℓ are computed with the subdivisions at level $\ell + 1$ by the following vector relationship $H^{c,\ell}(i) = \sum_{k=1}^{4} H^{c,\ell+1}(\jmath_k)$. We refer to the whole sequence of resolution-level histograms of image I as H_I. Like the feature counts in the original spatial-pyramid kernel, the class probability scores are amenable to this accumulation process. For a pair of score histograms H_{I_1} and H_{I_2} computed across all the initial classifiers c representing image I_1 and I_2, the spatial-pyramid matching kernel $\mathcal{K} = K(H_{I_1}, H_{I_2})$ is

$$\mathcal{K} = \sum_{\ell=0}^{L} \frac{1}{2^{L-\ell}} \left(\mathcal{I}(H^\ell_{I_1}, H^\ell_{I_2}) - \mathcal{I}(H^{\ell+1}_{I_1}, H^{\ell+1}_{I_2})\right) \tag{1}$$

where ℓ indexes the spatial resolution levels. $\mathcal{I}(H^\ell_{I_1}, H^\ell_{I_2})$ denotes the histogram intersection distance across all class channels and spatial cells of level ℓ. Note that for $\ell = L + 1$ this distance has zero value. The kernel $\mathcal{K}$ can handle different numbers of detections in each image. The similarity value that $\mathcal{K}$ represents is directly related to the number of descriptors and their classification score values. The weight associated with level ℓ is inversely proportional to the cell-width value; thus penalizing matches found in larger cells. These coarser-level matches are still used; they account for larger changes in position. Matches at the finest level are weighted the most, while still being robust to small changes in position. Our method then employs kernel $\mathcal{K}$ with the standard learning and prediction SVM algorithms. Experimental results described in Table 1 show the benefits of this image histogram descriptor with this type of kernel classifier, which outperformed the other stacked classifiers on the same histograms of feature combinations.

3. Stacked Training Set Creation and Learning

As indicated in the classification overview, our method is composed of two classification stages. The learning process thus requires three different procedures: (1) learning of the random trees classifiers $\{RT^c\}$ that will generate the local feature classification scores, (2) creation of the set of spatial histograms of scores $\mathcal{H}$ that constitutes the final classifier training set, and (3) learning of the final stacked spatial-pyramid classifier. This procedure is aimed at obtaining robust classifiers capable of handling all the variations present in the dataset while achieving high classification accuracy.

Random Trees Learning For each (detector, descriptor) pair c, a set of random trees RT^c is created from a training set $\mathcal{B}^c$. For each training image I with category y_I, every descriptor of set B^c_I of image I forms a training pair $(x^c_{I,j}, y_I)$. The training data $\mathcal{B}^c_\tau$ of size N of each tree τ is obtained through a bootstrap sampling procedure by drawing at random with replacement $N = |\mathcal{B}^c|$ descriptors with uniform probability. A set of Υ random trees is learned [4] from different $\mathcal{B}^c_\tau$ training sets, constrained for maximum tree depth and a minimum of 10 examples arriving at each leaf in the learning procedure. The parameter values for this learning step were determined experimentally. Figure 2 shows the relative insensitivity of the overall accuracy after 150 trees. In the tree learning procedure, every time a node is added, a subset of the attributes of the training examples of c (region descriptor and normalized region location) are randomly selected along with a threshold value as the the node splitting function. This combination of attributes allows the specialization of local classifiers in the first stage, which benefits the coupling of position and score accumulation as input of the stacked classifier. As part of the training process, for every tree τ, all the out-of-bag (OOB) training examples of τ (descriptors not used to train τ) are recorded. This information is then used in the next learning step to generate the training set for the stacked classifier.

Stacked Classifier Training Set Following the learning of the random trees classifiers, the training set for the stacked classifier is created. This set contains one spatial histogram per image. Let $\mathcal{H}^c$ be the set of labels and training histograms pairs obtained using single feature c and $\mathcal{H}$ be the set of label and training histogram pairs combining all the features from $\mathcal{Q}$. After the training of each RT^c classifier, the spatial histogram H^c_I of each image I in the training set is constructed using the same training descriptors from set $\mathcal{B}^c$ in a procedure similar to the one described in Section 2.2. The only difference in computing H^c_I is that for every descriptor $x^c_{I,j}$ of B^c_I being evaluated, its class-probability score $p^c_{I,j}$ is computed using only the votes from trees where this descriptor was an OOB element during training. Thus the values of $p^c_{I,j}$ are normalized by the number of OOB tree votes instead of the total number of trees. It is important to only use the trees where the example $x^c_{I,j}$ was never seen to model the behavior for unseen descriptors of a novel image being classified in the training set $\mathcal{H}^c$. The class label of image I is then assigned as the label of H^c_I. All the training pairs (H^c_I, y_I) are combined to create the stacked training set $\mathcal{H}^c$. The combined training set $\mathcal{H}$ is built by merging all the histograms H^c_I, of every training image I along the class-channel dimension.

Stacked Spatial Classifier Learning The parameters for the stacked spatial-pyramid kernel SVM learning algorithm are obtained by a logarithmic grid-search with a 5-fold cross-validation. For the multi-class experiments performed in this paper, a one-versus-all framework was employed. This learning stage of the stacked SVM classifier can be performed with a single-feature training set $\mathcal{H}^c$ or with a multiple-feature training set $\mathcal{H}$. The spatial-pyramid kernel $\mathcal{K}$ is used in the loss and discriminant functions of the SVM learning procedure.This kernel is well suited for the score histogram image representation H_I, because classification scores just like assignment counts [10], can be accumulated in the spatial pyramid.

4. Insect-Species Identification Experiments

In this section, we describe the series of experiments evaluating our method in the challenging *EPT29* dataset and report species and overall classification accuracy results. The *EPT29* dataset contains 4722 images divided into 29 species as indicated in the first column of Table 1. The specimens used to create this dataset were captured and identified by experts aiming to approximately emulate a biomonitoring sample for stream quality assessment. Note that some of these species have as many as ten times more images than the species with the fewest images. This type of imbalance tends to make identification tasks more challenging.

4.1. Random Trees Local Features

For the *EPT29* dataset experiments, our method has $C = 3$ initial RT^c classifiers. The (detector, descriptor) pairs c were selected to use regions with complementary position, shape, and orientation attributes; the descriptors for those regions were selected with that criteria as well. The pairs c employed are denoted as follows: dense grid of overlapping image patches with the HOG descriptor (*HOG*), salient points of high curvature with a beam angle and SIFT descriptors (*BAS+SIFT*), and the SIFT difference of Gaussians detector and descriptor (*SIFT*).

(1) The *HOG* [6] descriptors are obtained over a dense grid of 16×16 overlapping image patches, regardless of the image size. These patches overlap by a half to overcome

small changes in position. The HOG descriptor has proven really useful in detection tasks given the highly-redundant nature of its data. This grid-based feature can be applied even with the large changes in 3D position present in the data thanks to the orientation normalization process and the properties of the local features. (2) The *BAS+SIFT* descriptors are obtained on salient points of high curvature along the contour by using the IPAN [5] algorithm. For each salient point in the contour, a beam angle statistics (BAS) descriptor similar to [3] is generated and concatenated with a SIFT descriptor enhancing information by combining shape and appearance. To compute the BAS descriptor a set of lines, so-called beams, that originate from the anchor points connecting to each of the remaining salient points in the supporting region are employed. The angles between pairs of lines are calculated and accumulated with a weight inversely proportional to the perimeter distance. The BAS descriptor is the weighted histogram of these angles. The radius of the supporting region is selected adaptively to the length of the contour. A multiscale descriptor is generated by concatenating descriptors of different region size. (3) The *SIFT* features are detected and the descriptor generated using [11]. The region descriptors are scale invariant, isotropic, and invariant to rotation in the image plane.

4.2. Experimental Setup

All the different species-identification experiments are performed with a stratified 3-fold cross validation setup. Also, to make the results completely comparable, equal-sized random dataset partitions are the same in every fold. Given that for every specimen there were between 1–4 images obtained, the partitions are constrained to keeping all the images of any given specimen in a single one. In each of those images, the specimen is found in different 3D positions, orientations and poses. At every iteration of the experiment, two folds are used for training and one for testing. The results combining the predictions of the three testing procedures are reported. The values of the tree count Υ and the maximum tree depth parameters of every classifier c are determined experimentally. Figure 2 shows the behavior of the overall accuracy of *BAS+SIFT* features in relation to different parameter value combinations. The graph shows that accuracy increases flatten after 150 trees with a maximum depth of 25. For the other two pairs of $\mathcal{Q}$ the accuracy presents a similar behavior; thus the parameter values for the tree count Υ and maximum depth were set to 200 and 25 respectively for all three classifiers. For the spatial histogram H_I^c, the number of resolution levels L is set to 2 for a 4×4 grid at the finest level. The number of channels M is 29 constituting a 464-dimension image descriptor.

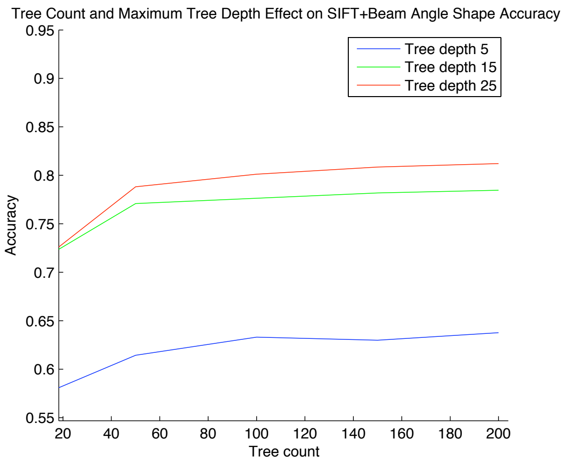

Figure 2. Overall accuracy graphs for the stacked spatial-pyramid kernel SVM with single feature pair BAS+SIFT showing the effect of the first-stage tree count Υ and maximum tree depth parameters.

4.3. Results

Table 1 reports the species and overall classification rates of the stacked spatial-pyramid kernel SVM method (Pyr), stacked random trees classifier (RTs 3Cmb) and RBF kernel SVM (RBF 3Cmb), and a single-level classifier with a χ^2 kernel SVM on a global HOG descriptor. The stacked spatial-pyramid kernel classifier is evaluated with every pair and with all combined. The other two stacked classifiers are only applied to the 3-feature combination (3Cmb). The table shows the effects of the number of images on the species accuracy, most of the species with the lowest maximum accuracy have the least number of images. The SVM classifier on a global HOG descriptor (χ^2 HOGgbl) is presented for comparison as a baseline, often used for state-of-the-art object-class detection. The benefit of stacking is clear by comparing the single stage HOG results to the local HOG stacked method. Results with χ^2 kernel are reported, instead of the standard linear kernel which were even lower. The results clearly indicate the beneficial coupling of the spatial histograms of scores with the spatial-pyramid kernel; the two non-pyramid stacked classifiers only had two of the highest species accuracies. Even the single feature spatial-pyramid classifier (Pyr SIFT) had an overall classification rate similar to the highest of the of the non-pyramid ones using the combined features. Finally, we note the large accuracy boost obtained by using feature combination (Pyr 3Cmb) with our stacked spatial-pyramid kernel method.

Discriminative Region Patterns Figure 3 shows representative high-score region patterns of the spatial-histogram channel m corresponding to the actual species of the displayed specimen. Each of the feature-type pairs c are shown

	Image	χ^2 HOGgbl	Pyr. HOGloc	Pyr. BAS	Pyr. SIFT	RTs 3Cmb	RBF 3Cmb	Pyr 3Cmb
Species	Count	[%]	[%]	[%]	[%]	[%]	[%]	[%]
Amphin	96	41.67	73.96	79.17	80.21	79.17	80.21	**85.42**
Asiop	292	95.55	95.21	95.89	95.21	96.92	95.21	**97.26**
Atops	254	75.20	86.61	89.37	87.80	87.40	86.22	**92.91**
Baets	251	72.11	85.26	83.67	79.68	81.27	86.06	**86.85**
Calib	299	60.20	71.24	73.58	79.93	77.93	75.59	**83.61**
Camel	287	77.00	83.62	83.28	86.76	82.58	85.71	**89.90**
Capni	130	42.31	60.77	77.69	79.23	75.38	76.15	**88.46**
Cerat	296	82.77	84.80	88.85	87.50	75.68	88.18	**90.88**
Cinyg	72	26.39	36.11	59.72	63.89	22.22	69.44	**79.17**
Cla	54	12.96	35.19	51.85	83.33	55.56	77.78	**87.04**
Culop	95	52.63	68.42	67.37	**81.05**	66.32	76.84	80.00
Drunl	42	23.81	61.90	78.57	78.57	59.52	66.67	**85.71**
Epeor	200	89.00	**93.50**	87.00	89.00	92.00	90.00	93.00
Fallc	224	36.61	44.64	54.46	62.95	46.43	**75.89**	64.73
Hlpsy	67	77.61	82.09	89.55	92.54	80.60	86.57	**95.52**
Isogn	229	78.60	86.46	87.77	**94.32**	81.22	93.45	93.89
Kat	48	58.33	41.67	64.58	**72.92**	29.17	70.83	**72.92**
Leucr	131	79.39	92.37	85.50	89.31	**96.18**	90.84	**96.18**
Limne	329	93.01	92.71	96.96	96.05	94.53	96.05	**98.18**
Lpdst	77	45.45	84.42	81.82	76.62	76.62	81.82	**87.01**
Lphlb	27	3.70	51.85	44.44	44.44	7.41	55.56	**59.26**
Meg	72	41.67	38.89	50.00	69.44	55.56	63.89	**77.78**
Mscap	132	58.33	67.42	73.48	75.00	68.94	80.30	**81.06**
Per	51	1.96	29.41	45.10	45.10	0.00	50.98	**52.94**
Plmpl	126	75.40	85.71	83.33	88.89	87.30	81.75	**92.06**
Siphl	150	60.00	78.00	**90.00**	88.67	85.33	85.33	**90.00**
Skw	292	63.01	71.92	78.42	77.74	61.99	81.51	**82.88**
Sol	129	86.82	87.60	87.60	94.57	92.25	93.02	**95.35**
Taenm	270	69.63	86.67	88.89	90.00	88.89	88.52	**91.48**
Total	4722	68.21	77.95	81.66	84.16	77.51	84.50	**88.06**

Table 1. Species accuracy results (All SVM classifiers except RTs). From left to right, first image count. Next, single-level global-HOG descriptor with χ^2 kernel. Following three, stacked spatial-pyramid kernel (Pyr) classifier with single feature pair c: local HOG patches (HOGloc), BAS+SIFT descriptors of salient curvature points (BAS), and SIFT detector and descriptor (SIFT). Last three, stacked classification using the 3-feature-types combination (3Cmb) of $\mathcal{Q}$: Random trees (RTs), RBF kernel (RBF), and spatial-pyramid kernel SVM (Pyr). Bold print indicates the highest accuracy for each species.

applied to individuals from the species that attained the highest species-identification accuracy of the single feature classifiers. The red parts of the insects are highly discriminative regions in the spatial histograms, which are positively correlated when evaluated with the spatial-pyramid kernel.

Evidence Histogram Results The spatial histograms H_I^c can also be computed by using descriptor probability scores $p_{I,j}^c$ based on leaf evidence [12] instead of simple leaf votes. We performed two experiments continuing using the experimental setup previously described; but with histograms computed by evidence scores from the *HOG* and the *BAS+SIFT* random tree classifiers RT^C. The overall accuracy results obtained are 77.47% and 80.62% respectively; which shows no significant difference with the accuracy obtained by using votes. The similar performance is probably explained by the large number of trees and the way the variables are aggregated in the stacked spatial-pyramid kernel SVM prediction procedure.

5. Conclusion

Our stacked spatial-pyramid kernel method enables the use of a discriminative stacking framework with the possibility of multiple-feature combinations while retaining spatial information. This method was applied on the large-scale *EPT29* dataset that emulates the insect specimen samples found in biological stream quality assessments. Automated insect species identification presents several chal-

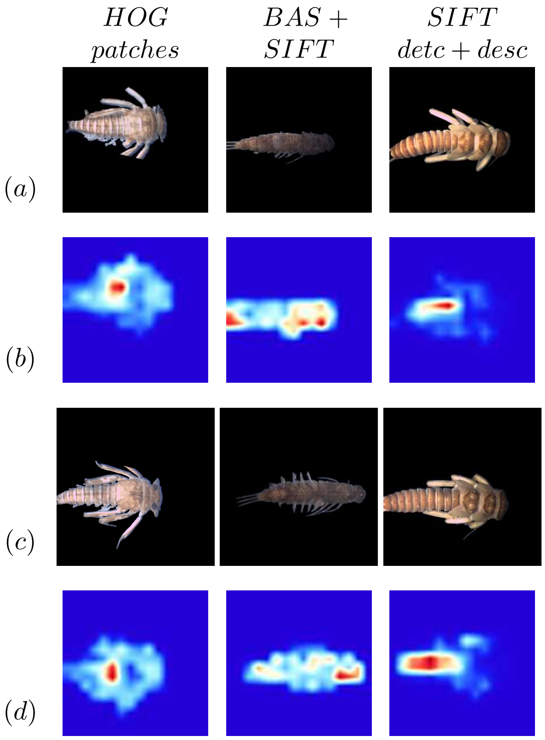

Figure 3. (a)(c) Example specimen images. (b)(c) Representation of a single channel m of histogram H_I^c. The channel m presented is of the actual species of the specimen. Red represents the highest scores, indicating some of the patterns of the most discriminative regions of that species. Each species is shown with the feature c that achieved the highest accuracy (Table 1). First, Epeor species with dense grid HOG patches, Siphl with salient curvature and BAS+SIFT descriptors, and Isogn with SIFT detector+descriptor.

lenges, such as small inter-species differences due to closely related species and significant intra-species variation due to changes in position, orientation, pose, and variations in developmental and degradation stages overcome by our method. The results obtained indicate that practical automated biomonitoring systems are possible. The spatial-pyramid kernel classifiers achieve higher classification rates than the random trees and RBF kernel classifiers, reaching a similar performance with just a single type of feature. A single-stage SVM classifier with a global HOG descriptor, which is often successfully applied in object-class detection tasks, was used as a baseline comparison of the two-level classification methods, and was significantly outperformed by our method. Finally, our method greatly benefited from combining multiple different features, with the experiments on the *EPT29* insect images attaining an overall classification boost of almost 4% over the single feature classifier with the highest accuracy.

References

[1] A. Abdullah, R. C. Veltkamp, and M. A. Wiering. Spatial pyramids and two-layer stacking svm classifiers for image categorization: a comparative study. In *IJCNN'09*, pages 1130–1137, Piscataway, NJ, USA, 2009. IEEE Press. 1, 2, 3

[2] T. Arbuckle et al. Biodiversity informatics in action: Identification and monitoring of bee species using ABIS. In *15th ISIEP*, pages 425–430, 2001. 2

[3] N. Arica and Y. Vural. BAS: a perceptual shape descriptor based on the beam angle statistics. *Pattern Recogn. Lett.*, 24(9-10):1627–1639, 2003. 3, 5

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 4

[5] D. Chetverikov. A simple and efficient algorithm for detection of high curvature points in planar curves. In N. Petkov and M. A. Westenberg, editors, *CAIP*, volume 2756 of *LNCS*, pages 746–753. Springer, 2003. 3, 5

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR '05*, pages 886–893, 2005. 2, 4

[7] Y. Gao, H. Song, X. Tian, and Y. Chen. Identification algorithm of winged insects based on hybrid moment invariants. pages 531 –534, jul. 2007. 2

[8] N. Larios et al. Automated insect identification through concatenated histograms of local appearance features. *Mach. Vision Appl.*, 19(2):105–123, 2008. 1, 2

[9] N. Larios, B. Soran, L. G. Shapiro, G. Martínez Muñoz, J. Lin, and T. G. Dietterich. Haar random forest features and svm spatial matching kernel for stonefly species identification. In *ICPR*, 2010. 1, 2

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06*, 2006. 1, 3, 4

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 3, 5

[12] G. Martínez Muñoz et al. Dictionary-free categorization of very similar objects via stacked evidence trees. In *CVPR' 09*, pages 549–556, 2009. 1, 2, 6

[13] M. Mayo and A. T. Watson. Automatic species identification of live moths. *Know.-Based Syst.*, 20(2):195–202, 2007. 2

[14] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *NIPS 19*, pages 985–992. MIT Press, Cambridge, MA, 2007. 1

[15] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR '08*, pages 1–8, 2008. 1

[16] S. Thiel and J. Ware. Determination of water quality in fresh water lakes. pages 662 –666, jul. 1995. 2

[17] C. Wen, D. E. Guyer, and W. Li. Automated insect classification with combined global and local features for orchard management. In *ASABE*, 2009. 2

[18] S. M. White, D. Marino, and S. Feiner. Designing a mobile user interface for automated species identification. In *SIGCHI '07*, pages 291–294, NY, USA, 2007. ACM. 2