A Hierarchical Object Recognition System Based on Multi-scale Principal Curvature Regions

Wei Zhang, Hongli Deng, Thomas G. Dietterich and Eric N. Mortensen
School of Electrical Engineering and Computer Science
Oregon State University, Corvallis, OR 97331, USA
E-mail: {zhangwe | deng | tgd | enm}@eecs.oregonstate.edu

Abstract

This paper proposes a new generic object recognition system based on multi-scale affine-invariant image regions. Image segments are obtained by a watershed transform of the principal curvature of a contrast enhanced image. Each region is described by an intensity-based statistical descriptor and a PCA-SIFT descriptor. The spatial relations between regions are represented by a cluster-index distribution histogram. With these new descriptors, we develop a hierarchical object recognition system which uses an improved boosting feature selection method [9] to construct layer classifiers by automatically selecting the most discriminative features in each layer. All layer classifiers are then combined to give the final classification. This system is tested on various object recognition problems. Experimental results show that the new hierarchical system outperforms the comparable solutions on most of the datasets tested.

1. Introduction

The description of object classes is a crucial issue in the design of object recognition systems. Previous description methods include single-scale fragment-based [12] and interest-region approaches [1,3,5,8,9]. While fragment or part features are usually very informative for object categories, they can be too class-specific and are not transform invariant. Interest regions are more generic and more robust to occlusion and transformations, but they are too local and often noisy. Probabilistic constellation models [5] and clustering-based methods [3] have been proposed to recognize image categories based on these fragments or interest regions.

Instead of describing objects at a single scale, other methods represent object information at multiple scales [2,4,10]. These descriptions are biologically motivated — the human visual system selects and combines both coarse (global) and detailed (local) object features for recognition. Shokoufandeh et al. [10] use saliency map graphs to capture the salient image structure using multi-scale wavelet transforms. Epshtein and Ullman [4] propose feature hierarchies based on mutual information feature selection and parameter adaptation. The work of Bouchard and Triggs [2] model each object as a hierarchy of parts and subparts with partial transformations (translation and scale transformations) that softly relate the parts and sub-trees to their parents. But there is a common weakness existing in these hierarchical object descriptions: all these descriptions are highly concrete models (trees or graphs). Applying these types of descriptions to classification requires graph matching [10] or model instantiation [2,4] algorithms.

In this paper, we propose a new generic object description which characterizes the global and local features of an object class based on multi-scale principal curvature regions. This new object description method is introduced in Section 2. Given these multi-scale descriptors, Section 3 introduces and details a hierarchical object recognition system using an improved boosting feature selection method. Finally, experimental results and conclusions are given in Sections 4 and 5, respectively.

2. Multi-scale principal curvature regions

2.1. Decoloring

To facilitate computation of feature descriptors, we convert color images to intensity images while preserving contrast between regions. We employ the decolorization algorithm proposed by Grundland and Dodgson [6] to do color contrast enhancement when converting color images to grayscale images. This algorithm enhances contrast in a meaningful way by adjusting luminance to reflect chromatic differences.

2.2. Multi-scale principal curvature

We adapt the curvilinear structures detector of Steger [11] to generate structural object regions defined by the watershed of the image's principal curvature. It has been our experience that using the principal curvature produces fairly stable regions that can be detected over a range of viewpoints, scales, and appearance changes. Further, these regions seem more characteristic of object classes compared with local corners and blobs.

The local shape characteristics of an image, viewed as a surface, can be described by the Hessian matrix

$$\mathbf{H}(\mathbf{x},\sigma_D)=\begin{bmatrix} I_{xx}(\mathbf{x},\sigma_D) & I_{xy}(\mathbf{x},\sigma_D) \\ I_{xy}(\mathbf{x},\sigma_D) & I_{yy}(\mathbf{x},\sigma_D) \end{bmatrix} \qquad (1)$$

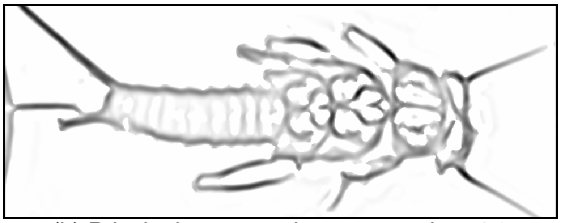
where $I_{xx}(\mathbf{x},\sigma_D)$, $I_{xy}(\mathbf{x},\sigma_D)$, and $I_{yy}(\mathbf{x},\sigma_D)$ are the second-order derivatives of the image as computed by convolving the image with the appropriate second-derivative of a Gaussian with scale σ_D. For every pixel, the eigenvalues, (λ_1, λ_2), of (1) are proportional to the local principal curvatures while the corresponding eigenvectors, $\mathbf{v}_1$ and $\mathbf{v}_2$, specify the directions of principal curvature. Each pixel of the principal curvature image is simply the largest eigenvalue of $\mathbf{H}(\mathbf{x}, \sigma_D)$. To reduce the impact of noise, we suppress all principal curvature pixels (i.e., assign them to zero) that are below a threshold τ_0. We then apply a watershed transform to the "cleaned" principal curvature image using the immersion simulation method proposed by Vincent and Soille [13]. Finally, each watershed region is approximated with an ellipse having the same second moment.

To extract a multi-scale object description that integrates both global and local information of object class, we apply the region detection algorithm at various scales. Figure 1 shows examples of multi-scale principal curvature regions. We see that a larger scale produces fewer and global regions while a smaller scale results in more local features.

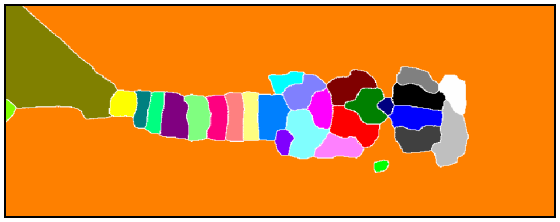
2.3. Principal curvature region descriptions

To more comprehensively describe each region, we employ both statistical measurements of region intensities and PCA-SIFT [7] features. The statistical feature combines coefficient of variation, skewness, kurtosis, and moment invariants to form a 9-dimensional feature vector for each region. The PCA-SIFT features are 36-dimensional and have been demonstrated to be more compact and distinctive than SIFT [7].

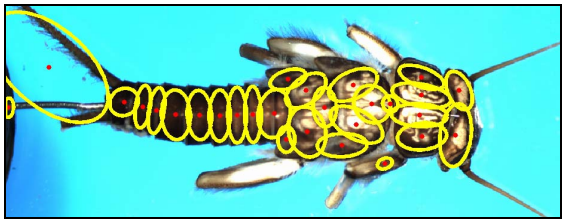
(a) Original image



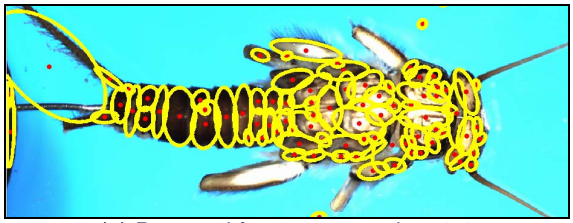
(b) Principal curvature image at scale = 4.0



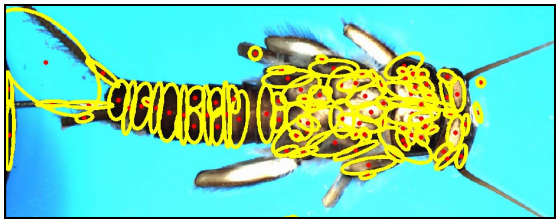
(c) Segmented regions at scale = 4.0



(d) Detected features at scale = 4.0



(e) Detected features at scale = 2.0



(f) Detected features at scale = 1.0

Figure 1. Multi-scale region detections.

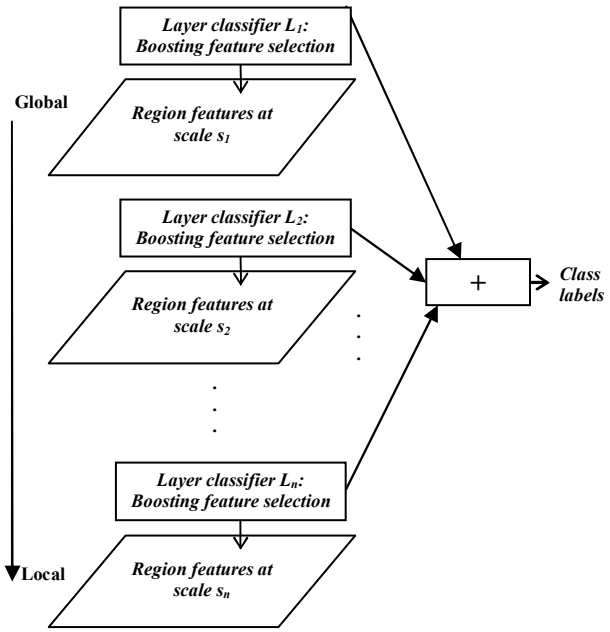


Figure 2. Hierarchical object recognition system.

In addition, we characterize the spatial configuration of the regions with bins-based cluster index distribution histograms. The construction of spatial relation features involves three steps. First, we cluster the PCA-SIFT features from the positive training images using E-M to fit a Gaussian mixture model with $C = 16$ clusters. Second, for each region in the training and testing images, we compute the index of the Gaussian cluster most likely to have generated its PCA-SIFT vector. And third, we discretize the distances and directions between regions into $M = 36$ bins with 12 directions and 3 distance ranges. The sizes of the bins are fixed relative to the image sizes. Thus, the spatial configuration of regions in each image is described by a histogram R composed of $D = C \times M \times C = 16 \times 36 \times 16 = 9216$ feature elements. An element $R_{i,m,j}$ in R records the number of times a region with cluster index j falls into bin m with center region index i.

3. Hierarchical object recognition system

Using our new object descriptions, we designed a hierarchical object recognition system which uses multi-scale image analysis to do classification. This system is illustrated in Figure 2. From the top layer to the bottom, we train layer classifiers $L_1,...,L_n$ based on the region features obtained at scales $s_1,...,s_n$, which are in decreasing order (global to local). We then combine the outputs of layer classifiers to predict the class labels of new images.

3.1. Layer classifier

Using our new description method above, object images are described by normal feature vectors of three types (intensity statistical features, PCA-SIFT, and spatial relation features) instead of concrete models. This permits standard classification algorithms to be employed as layer classifiers. According to our experiments, we noticed that for most of the image sets, only a small portion of the image features are useful for classification. So we employ and improve the boosting feature selection algorithm proposed by Opelt et al. [9] that searches among all the available features and automatically selects the most stable and discriminative ones to form the final classifier.

The layer classifiers are learned using the AdaBoost algorithm which maintains a weight for each training image. In iteration t of AdaBoost, all the unselected feature vectors of the training images are evaluated based on the current image weights to find the most discriminative feature.

We evaluate the statistical intensity features and the PCA-SIFT features in the same way as Opelt et al. [9]. The stability and discriminating power of a feature vector v_f is evaluated in three steps. First, calculate the distance from v_f to each of the training images. This is done by finding the minimum distance between v_f and all the feature vectors of the same type in the training image. We use the Mahalanobis distance metric for the statistical intensity feature and the Euclidean distance for PCA-SIFT. Second, sort the training images into ascending order according to their distances to v_f. Third, we apply the scanline algorithm [9] to the sorted distance array to determine a threshold θ_f that maximizes the weighted accuracy of using v_f as a weak classifier. The maximal weighted sum is adopted as the evaluation of v_f.

Evaluating the spatial relation features is simpler because there is no need to calculate the feature-to-image distances. The training images are directly sorted according to their spatial relation feature values. More specifically, all the spatial relation features of K training images are assembled into a $D \times K$ matrix A (where D is the dimension of the spatial configuration histogram). Then for each row of A, training images are sorted by decreasing order of their corresponding feature values. Finally, the scanline algorithm scans the sorted array and outputs the optimal threshold and the maximal weighted sum evaluation for the row, which indicates the significance of the specific spatial configuration for classification.

A perfect feature should have all of the positive images (+1) sorted before all the negative images (–1) so that the feature vector gives a weak classifier that is perfectly discriminative. The feature and threshold {v*,

$\theta^*\}$ which has maximal score among all the available feature vectors is selected as the weak classifier for iteration t. We construct T weak classifiers for each layer. All these T weak classifiers are then combined into a strong classifier (called the layer classifier) using standard AdaBoost. The output of a strong classifier L_i is given by

$$y_i = \sum_{t=1}^{T} (\ln \beta_{i,t}) h_{i,t}(I) \qquad (2)$$

with

$$\beta_{i,t} = \sqrt{\frac{1-\varepsilon_{i,t}}{\varepsilon_{i,t}}} \qquad (3)$$

where $h_{i,t}(I)$ represents the output of the t_{th} weak classifier of layer classifier L_i. $\varepsilon_{i,t}$ is the weighted classification error rate of the t_{th} weak classifier computed based on the AdaBoost weights.

For presence/absence 2-class object recognition problems, it is not plausible to use background features to recognize object examples. So we modified the original algorithm in [9] to select only among the features from positive images.

3.2. Final classification

The final result of the hierarchical system is simply the sign of the sum of the outputs of layer classifiers, which is given by:

$$Y = sign(\sum_{i=1}^{n} y_i) \qquad (4)$$

In our tentative experiments, we also tried to set weights for layer classifiers, and use the Voted Perceptron algorithm to adapt the weights to minimize the classification error on training images, but it overfits the data and the performance degrades.

4. Experimental results

We did experiments on various 2-class object recognition image sets in order to test the performance of our system. We employed a four-layer system with scales {4.0, 3.0, 2.0, 1.0} and the number of boosting iterations T = 100. The system is tested on six object classes in the Caltech dataset[1]: airplanes (1074), cars (rear) (526), cars (side) (123), faces (450), leaves (186) and motorbikes (826). The background set in Caltech contains 451 images. We also tested on a stonefly larva set containing 70 Doroneuria images (positive) and 57 Hesperoperla images (negative). Examples of Caltech images and stonefly images are shown in Figure 3. Half of the images in each set are used for training, and the rest are held out for testing. Recognition performance is evaluated by ROC equal error rates.

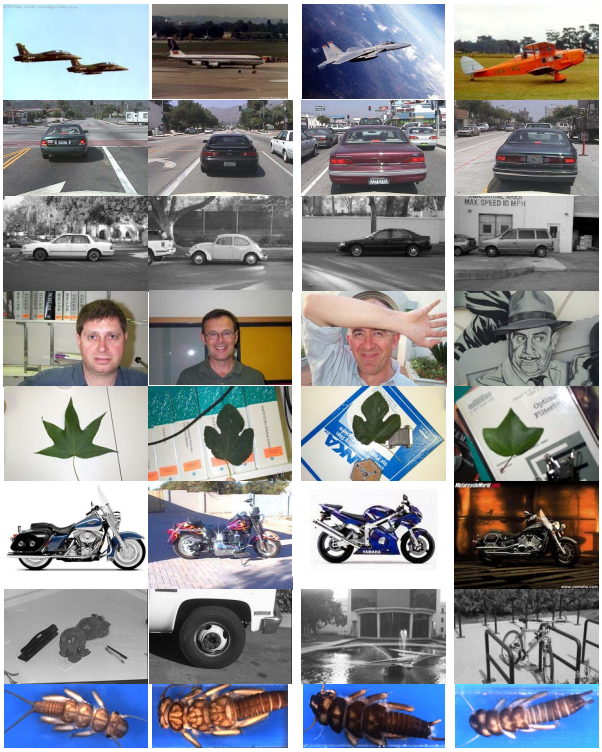


Figure 3. Sample images from Caltech and stonefly larva dataset with rows corresponding to: airplanes, cars (rear), cars (side), faces, leaves, motorbikes, Caltech background, and Doroneuria (left two images) and Hesperoperla (right two images).

The hierarchical system based on the new descriptions is tested on these datasets and compared with the constellation model of Fergus et al. [5] and the boosting feature selection approach by Opelt et al. [9]. The results are summarized in Table 1. The comparison indicates that our hierarchical object recognition system outperforms the other methods on most of the comparable datasets.

In order to test the value of our hierarchical structure, we compared the equal error rates of the whole 4-layer system (denoted as 4-layer with spatial) to the best single layer classifier (1-layer). The results are summarized in the second and third columns of Table 2. In the forth column of Table 2, we show the performance of the 4-layer system without spatial relation (4-layer without spatial) to test the utility of the spatial configuration descriptor.

We noticed that on all these datasets, there are significant gaps between the performance of the multi-

[1] http://www.vision.caltech.edu/feifeili/Datasets.htm

layer system and that of the best one-layer classifier. This demonstrates that the multi-scale object description is more generic and informative for object classes than single scale description.

On most of the datasets, spatial relation features improve the performance of the system, thus supporting our claim that spatial configurations of detected regions are also valuable cues for recognition.

Table 1. ROC equal error rates of our approach and other approaches.

Dataset	Ours	Fergus [5]	Opelt [9]
Airplane	90.6	90.2	88.9
Cars(rear)	94.3	90.3	/
Cars(side)	83.6	88.5	83.0
Faces	98.8	96.4	93.5
Leaves	97.5	/	/
Motorbikes	94.3	92.5	92.2
Stoneflies	88.6	/	/

Table 2. ROC equal error rates of the 4-layer classifier with spatial relation features compared to 1-layer classifier and 4-layer classifier without spatial relation features.

Dataset	4-layer with spatial	1-layer	4-layer without spatial
Airplanes	90.6	89.0	90.0
Cars(rear)	94.3	91.0	89.2
Cars(side)	83.6	81.6	80.3
Faces	98.8	97.2	98.8
Leaves	97.5	96.0	97.3
Motorbikes	94.3	92.0	93.5
Stoneflies	88.6	80.0	82.9

5. Conclusion and future work

In this paper, we propose a novel object description based on multi-scale principal curvature regions. This description is invariant to rotation and view transformations and robust to scale changes. The texture and geometric information of detected regions are represented by their intensity-based and spatial relation features respectively. A generic hierarchical object recognition system using boosting feature selection is developed and outperforms two other approaches on various object classes.

There are two future directions we wish to investigate. One is to further improve the robustness of the principal curvature region detector. The other is incorporating inter-layer spatial relations into the hierarchical system to further exploit the spatial constraints.

6. Acknowledgements

This work is supported by National Science Foundation grant number 0326052 entitled "ITR: Pattern Recognition for Ecological Science and Environmental Monitoring". We would like to thank Caltech Vision Lab and Yan Ke for providing test dataset and source code for descriptor.

7. References

[1] S. Agarwal, A. Awan and D. Roth, "Learning to Detect Objects in Images via a Sparse, Part-Based Representation," *IEEE PAMI*, 26(11):1475-1490, 2004.

[2] G. Bouchard and Bill Triggs, "Hierarchical part-based visual object categorization," *CVPR*, 2005.

[3] G. Dorko and C. Schmid, "Object Class Recognition Using Discriminative Local Features," submitted to *PAMI*, 2004.

[4] B. Epshtein and S. Ullman. "Feature Hierarchies for Object Classification," *ICCV*, 2005.

[5] R. Fergus, P. Perona, and A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning," *CVPR*, 2003.

[6] M. Grundland and N. A. Dodgson, "The Decolorize Algorithm for Contrast Enhancing, Color to Grayscale Conversion," *Technical Report UCAM-CL-TR-649,* University of Cambridge.

[7] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," *CVPR*, 2004.

[8] J. Matas, O. Chum, M. Urban, T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," *BMVC*, 2002.

[9] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. "Weak Hypotheses and Boosting for Generic Object Detection and Recognition," *ECCV*, 2004.

[10] A. Shokoufandeh, I. Marsic, and S. J. Dickinson, "View-Based Object Recognition Using Saliency Maps," *Image and Vision Computing*, 17(5–6):445–460 (1999)8.

[11] C. Steger, "An Unbiased Detector of Curvilinear Structures," in *IEEE PAMI*, 20(2):113-125, 1998.

[12] S. Ullman, E. Sali and M. Vidal-Naquet, "A Fragment-Based Approach to Object Representation and Classification," in *International Workshop on Visual Form*, Berlin: Springer, 85-100, 2001.

[13] L. Vincent and P. Soille, "Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations," *IEEE PAMI*, 13(6):583-598, 1991.