Proceedings of the Sixteenth International Conference on Pattern Recognition, 2002.

Fusion of Global and Local Information for Object Detection

Ashutosh Garg[†], Shivani Agarwal[§], Thomas S. Huang[†]
[†]Beckman Institute, [§]Department of Computer Science
University of Illinois, Urbana, IL 61801
{ashutosh,sagarwal,t-huang1}@uiuc.edu

Abstract

This paper presents a framework for fusing together global and local information in images to form a powerful object detection system. We begin by describing two detection algorithms. The first algorithm uses independent component analysis (ICA) to derive an image representation that captures global information in the input data. The second algorithm uses a part-based representation that relies on local properties of the data. The strengths of the two detection algorithms are then combined to form a more powerful detector. The approach is evaluated on a database of real-world images containing side views of cars. The combined detector gives distinctly superior performance than each of the individual detectors, achieving a high detection accuracy of 94% on this difficult test set.

1 Introduction

Object detection in images is an important problem that has recently gained a lot of attention in the vision community. The main challenge in object detection arises from the wide range of variations across different imaging conditions and across different objects in the object category of interest. To be successful, any approach to the problem must be able to generalize over these variations.

Most approaches that have been proposed for the problem rely on some underlying feature representation to facilitate this generalization. The feature representation can be either the raw pixel-based representation, or a higher level representation obtained by applying some transformation to the raw image data. Different representations capture different kinds of information in the data. In particular, some representations capture global information in images, while others capture local properties. Although both approaches work well to a certain extent, each is limited by the fact that it ignores other information that may also be important.

In this paper, we present an approach that allows both global and local information to be fused together. In particular, we describe two detection algorithms, one that uses global information in images and another that relies on local information, and show how these two kinds of information can be merged together to form a more powerful detection system. As a test bed for our approach, we choose a difficult test set of real-world images that contain side views of cars against varied natural backgrounds. Our experiments on this test set show that combining information from the two detectors outperforms each of them individually.

We present our first detection algorithm, which captures global properties of the data using independent component analysis (ICA), in Section 2. Section 3 describes the second detection algorithm, which uses a part-based representation to capture local properties of the data. In Section 4 we present our approach for fusing the two kinds of information. Section 5 presents our experimental results, followed by conclusions in Section 6.

2 ICA-Based Approach: Global Information

In this approach, we use ICA to extract a set of independent components from some sample object images. These independent components are used as a basis to form a representation for images. Boosting is then used to learn an object/non-object classifier based on this representation, and the learned classifier is used to develop a detector for the object class.

2.1 Representation using ICA

ICA can be viewed as an extension of Principal Component Analysis (PCA). PCA relies only on the second order properties of the data, thus ignoring much of the information that may be contained in higher order relationships among the image pixels. ICA separates higher order moments of the input in addition to second order moments.

The idea behind using ICA is to find a set of statistically independent "source" images for a set of object images. This is done using the following model:

$$X = AS \tag{1}$$

where X represents the observed images, S the (unknown) independent sources, and A the (unknown) mixing matrix. The procedure described in [2] is used to find the unmixing matrix W. This is then used to obtain the unmixed sources U, which are an estimate of the original sources S under some permutation (details are omitted due to lack of space):

$$U = WX \tag{2}$$

In our experiments, we applied this technique to 200 car images, where each image is of size 100×40. Some of the independent components obtained are shown in Figure 1. These form a set of independent basis images, and can be

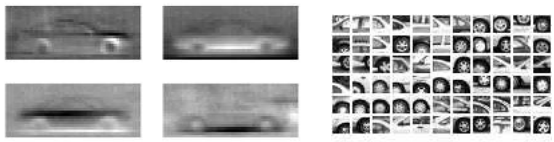

Figure 1. Left: Examples of the independent components obtained using the ICA approach; this representation captures global information. **Right:** Examples of the parts obtained using the part-based approach; this representation captures local information.

viewed as providing a set of statistically independent image features. Each image is represented in terms of this basis as a feature vector containing the normalized dot products of the image with each of the independent components. Since this representation requires operating on the entire image, it can be viewed as capturing global information in the image.

2.2 Classification using Boosting

Boosting is a technique for combining weak learners to form a strong classifier. It has been used successfully in various problems in vision, including image retrieval [6], face detection [7] and image segmentation [5]. In all these approaches, an image is represented as a set of some features. Each feature (with an associated threshold) is then treated as a base classifier, and boosting is performed to learn a strong classifier as an ensemble of these base classifiers.

In our method, each 100×40 training image (and later each 100×40 window in the test images) is represented as a feature vector obtained from the independent component basis (as described above). Boosting is then performed over these features. We use the AdaBoost algorithm described in [4]. At each iteration, we model each feature as a mixture of two gaussians, with a gaussian corresponding to each class (object vs. non-object). Based on the current weight over the data, the optimal threshold is computed; the feature and this learned threshold are then used to classify the data. The feature that gives minimum classification error on the training set, along with the corresponding threshold, is chosen is as the base classifier for that particular iteration. A final classifier h_1 is then learned as a combination of these base classifiers, weighted and thresholded appropriately:

$$h_1(x) = \text{sgn}\left(\left[\sum_i w_i \text{sgn}(f_i(x) - \theta_i)\right] - \theta\right) \quad (3)$$

where each f_i denotes a feature of the input x, θ_i the corresponding learned threshold, w_i the weight assigned to this feature by the boosting algorithm, and θ the overall threshold. (Again, details are omitted due to lack of space; the approach is similar to that in [5].) In addition to the binary classification given by the thresholded combination, the absolute value (without the threshold) can be used to give an activation or confidence value:

$$\alpha_1(x) = \sum_i w_i \text{sgn}(f_i(x) - \theta_i) \quad (4)$$

The activation values are used to form a classifier activation map that allows the learned classifier to be used as an effective detector; this step is discussed in Section 4.1.

3 Part-Based Approach: Local Information

In this section we describe an object detection algorithm that uses a part-based representation. Only a brief overview is given here; a detailed description can be found in [1].

3.1 Representation using Parts

A vocabulary of parts is first constructed automatically from a set of sample object images. This is done by using an interest operator to select points in the images based on local signal behavior; small image patches are then extracted around these interest points to form the part vocabulary. Some of the parts obtained are shown in Figure 1.

This part vocabulary is then used to represent images. Given a new image, we determine which of the vocabulary parts are present in it. This is done by finding interest points in the image to focus attention on the interesting regions, and then comparing local patches around these points to parts in the vocabulary. Vocabulary parts that are sufficiently similar to such interest patches are considered to be present in the image. The parts thus detected in the image, together with spatial relations among them, are used as binary features to form a representation for the image. Since this representation requires only local operators, it can be viewed as capturing local information in the image.

3.2 Classification using SNoW

Based on the above part-based feature representation, a classifier to classify 100×40 images as object or non-object is learned using the SNoW learning architecture [3]. SNoW is a feature-efficient variation of the winnow learning algorithm, and is therefore useful in this case as the potential number of part and relation features is very large. SNoW learns a linear threshold function for each class in the classification task (in this case, object and non-object). Given an input x, the classifier h_2 predicts the class according to the following function:

$$h_2(x) = \text{sgn}\left(\left[\sum_i (w_i^+ - w_i^-) f_i(x)\right] + (\theta^+ - \theta^-)\right) \quad (5)$$

where w_i^+ and θ^+ are the weights and threshold of the function corresponding to the positive (object) class, and w_i^- and θ^- those corresponding to the negative (non-object) class. (Note: the features f_i here are different from Section 2.2; they are binary features indicating the presence or absence of a vocabulary part, or of a spatial relation between two such parts.) When applied to an input vector, the linear threshold function corresponding to the positive class can be used to give an activation or confidence value:

$$\alpha_2(x) = \begin{cases} \sum_i w_i^+ f_i(x) - \theta^+ & \text{if } h_2(x) = +1 \\ 0 & \text{if } h_2(x) = -1 \end{cases} \quad (6)$$

Again, these activations are used to form an activation map for detection; this step is described in Section 4.1.

4 Fusion of Global and Local Information

The detection algorithm described in Section 2 uses primarily global information in images, whereas the algorithm of Section 3 relies primarily on local information. This section presents an approach for fusing together the global information from the first method and the local information from the second to form a more powerful detection system. We make use of the confidence of each scheme which is then used to make the combined decision. The activations of each of these methods are treated as their confidence in the classifier output. We first describe how the activations produced by each classifier are used to form a classifier activation map for detection. We then discuss how the activation maps produced from the two individual classifiers are merged together to combine information from both.

4.1 Classifier Activation Map

Given a classifier h that can classify an image as positive (object) or negative (non-object) based on some activation value α it produces, detection in an image proceeds by shifting a 100×40 window over the image and applying the classifier to each such window w. The activation values $\alpha(w)$ produced by the classifier at the various window locations yield a classifier activation map [1]. This map has high values at locations where the classifier has high confidence in its positive classification. The algorithm described in [1] is used to analyze this map for activation peaks, giving accurate localization of objects and preventing multiple detections correponding to a single object in the image.

4.2 Fusion of Classifier Activation Maps

Given a test image, the ICA-based classifier h_1 produces an activation map based on its activations α_1, while the part-based classifier h_2 produces an activation map based on its activations α_2.[1] Each map captures information based on the corresponding classifier, which in turn captures information based on the representation it uses. Each of these maps can be used to perform detection as described above.

To combine information from the two detectors, the activations α_1 and α_2 from the two classifiers are combined via a function γ to obtain a new activation α_{fused}. This allows the two original activation maps to be merged into a new activation map, in which the activation for any image window w is given by

$$\alpha_{\text{fused}}(w) = \gamma(\alpha_1(w), \alpha_2(w)) \tag{7}$$

Different choices for the function γ above combine the two kinds of information in different ways. To find an effective combination, we learn γ automatically from the training images. Each training image x is represented by the 2-element vector $(\alpha_1(x), \alpha_2(x))$. These training vectors, along with the corresponding positive (object) or negative (non-object) labels of the images, are then fed as input to a supervised learning algorithm to learn a function that gives good classification accuracy on the training set.

In our implementation, γ is learned as a function of the individual activations by a simple perceptron. Experiments were performed with both linear and quadratic functions. (Note that a linear learning algorithm can be used to learn a higher degree polynomial by adding higher order terms and then learning the coefficients of these terms as a linear function in the new space.) We found that a quadratic function of the individual activations gives high classification accuracy on the training set; this is used in the final implementation.

The new activation map thus obtained contains information from both detectors. The same inference algorithm used on the earlier activation maps can now be applied to this new map to form a better detection hypothesis. The approach described here can be used to combine information from any set of detectors that rely on classifiers which can give activation values to produce activation maps; by learning a simple function over these activations, the individual maps can be merged into a new map that allows better inference for object detection.

5 Experimental Results

We evaluated our approach on the car image database used in [1]. The database contains 1000 training images (500 positive and 500 negative), and 170 test images containing 200 cars in all. The images in this database are all natural; they are taken from different sources and include images with occlusion and cluttered backgrounds. The evaluation scheme used for evaluating detections as correct or false is the same as that used by [1].

Figure 2 shows the results of the three approaches on this test set. The performance of our ICA-based approach is comparable to that of the part-based approach. However, when the strengths of the two approaches are combined using our fusion method, the performance of the resulting detector surpasses each of the individual detectors.

For each of the three methods, the point at which that method gives highest recall (together with highest precision at that recall) is shown in Table 1. The combined detector gives distinctly higher accuracy than the individual detectors in terms of both recall and precision. Figure 3 shows samples of the output of our detector.

6 Conclusions

We argued that both global and local information are important for the object detection task. We described two detection algorithms - one that uses global information via

[1] In the implementation, the maps were obtained by shifting a 100×40 window in steps of 4 pixels in the horizontal direction and 2 pixels in the vertical direction.

Figure 3. Examples of test images on which our combined detector achieved perfect detection results. Note that the windows are drawn at the *exact* locations output by the detector. Multiple detections corresponding to a single object are not allowed.

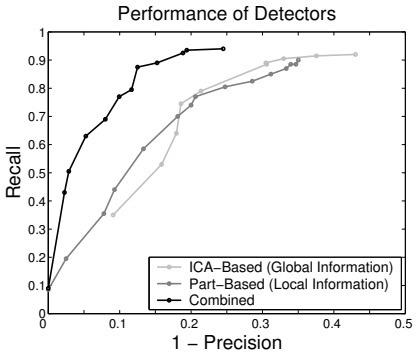

Figure 2. Performance of the different methods on the car image test set containing 200 cars. *Recall* measures the number of correct detections relative to the total number of positives (here, 200) in the database; (1 – *precision*) measures the number of false detections relative to the total number of detections made by the system. Different points on the curve are obtained by varying an activation threshold parameter in the classifier activation map. The combined detector achieves distinctly superior performance than each of the two individual detectors. (Important: note that the X-axis range is [0, 0.5] for clarity.)

an ICA-based representation, and another that uses local information via a part-based representation - and showed how a function that combines the two kinds of information in an effective way can automatically be learned from the training images. We showed that combining the two kinds of information gives distinctly superior performance than the individual detectors. The combined detection algorithm achieves a high detection accuracy of 94% on a difficult test set of real-world images of cars.

Our approach shows that a powerful object detection system can be built by fusing different kinds of information in images. Our framework can easily be extended to combine information from several different detectors. The advantage of this approach is that it can be used to look at several different aspects of the input data in parallel, and the outputs of the individual detectors can then be combined in a simple step to form a highly accurate detection system.

Detection method	*No. of correct detections*, N	*Recall* $\frac{N}{200}$	*No. of false detections*, M	*Precision* $\frac{N}{N+M}$
ICA-based	184	92.0%	139	56.97%
Part-based	180	90.0%	97	64.98%
Combined	**188**	**94.0%**	**61**	**75.50%**

Table 1. For each detection method, the point at which highest recall is achieved (and highest precision at that recall) is shown. The combined detector gives higher accuracy than the individual detectors in terms of both recall and precision.

Acknowledgments: We are grateful to Dan Roth for discussions related to this work. This research was supported by NSF grants ITR-IIS-0085836, ITR-IIS-0085980 and IIS-9984168, and an IBM fellowship to Ashutosh Garg.

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *Proceedings of the Seventh European Conference on Computer Vision*, 2002. To appear.

[2] M. S. Bartlett, H. M. Lades, and T. J. Sejnowski. Independent component representations for face recognition. In *Proceedings of the SPIE Conference on Human Vision and Electronic Imaging III*, volume 3299, pages 528–539, 1998.

[3] A. J. Carlson, C. Cumby, J. Rosen, and D. Roth. The SNoW learning architecture. Technical Report UIUCDCS-R-99-2101, UIUC Computer Science Dept, May 1999.

[4] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 148–156, 1996.

[5] V. Pavlovic and A. Garg. Efficient detection of objects and attributes using boosting. In *Technical Sketches, IEEE Conference on Computer Vision and Pattern Recognition*, 2001. To appear.

[6] K. Tieu and P. Viola. Boosting image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 228–235, 2000.

[7] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001.