Object Detection by Estimating and Combining High-Level Features

Geoffrey Levine and Gerald DeJong

Department of Computer Science
University of Illinois at Champaign-Urbana
Urbana, IL 61801
levine@cs.uiuc.edu, dejong@cs.uiuc.edu

Abstract. Many successful object detection systems characterize object classes with a statistical profile over a large number of local features. We present an enhancement to this method that learns to assemble local features into features that capture more global properties such as body shape and color distribution. The system then learns to combine these estimated global features to improve object detection accuracy. In our approach, each candidate object detection from an off-the-shelf gradient-based detection system is transformed into a conditional random field. This CRF is used to extract a most likely object silhouette, which is then processed into features based on color and shape. Finally, we show that on the difficult Pascal VOC 2007 data set, detection rates can be improved by combining these global features with the local features from a state-of-the-art gradient based approach.

1 Introduction

Recently, the field of computer vision has taken great strides in the area of object detection. Many of today's top performing systems perform recognition based on a constellation of local gradient features [1,2]. These systems have performed well on many of the latest object detection datasets [3,4]. However, the performance of these systems varies dramatically from object to object. For example, in the recent PASCAL VOC 2008 detection challenge [4], top performing local feature approaches performed very well on object classes with predictable structures such as bicycle, car and train. On the other hand, objects that are highly deformable and/or viewed from a diverse set of perspectives, such as bird, dog, and plant, were detected less than half as accurately.

In Figure 1 we show several false detections of one such approach [1]. This approach models object classes with a spatial histogram of gradients. Indeed, it is possible to see how in each case the local gradient features could confuse the detector. Still, to a human, these cases are not problematic as there are clear cues that the objects in question are not present.

In this paper, we present a more directed approach, in which high-level color and shape consistency features are estimated and utilized for object detection. We demonstrate that by combining these high-level features with confidence values from a state-of-the art object detection system, we are able to improve detection accuracy on the difficult subclass of animals from the PASCAL VOC 2007 dataset [3].

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 161–169, 2009.

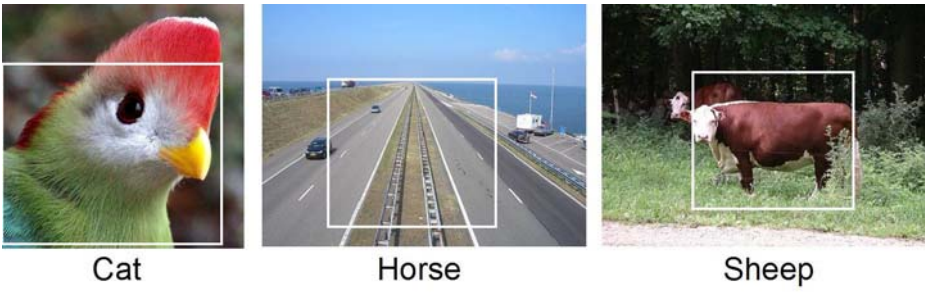

Fig. 1. False Positives (at equal error rate) of Felzenszwalb et al. [1]. The local features found by the gradient-based detector are consistent with the proposed class. However, in these cases, shape and color consistency can provide a strong cue that the object of interest is not present.

2 Approach

Our approach is illustrated in Figure 2. A black-box gradient-based object detector [1] is used to hypothesize object locations. We extract an image subwindow around each candidate detection, and construct a conditional random field probability model to estimate the most likely silhouette corresponding to each detection. These silhouettes are processed for global object shape and color features, which are amended to features available from the gradient-based detector, and a hybrid detector is trained to more accurately score candidate object detections.

Conditional Random Field. Conditional random fields [5,6,7], a probabilistic framework for classifying structured inputs, represent the conditional probability of a labeling, $\mathbf{Y}$, given structured data, $\mathbf{X}$, in the form of an undirected graphical model. The probability of a label sequence given observed data is equal to a normalized potential function:

$$P(\mathbf{Y}|\mathbf{X}) = \frac{e^{\psi(\mathbf{Y},\mathbf{X})}}{Z(\mathbf{X})} . \tag{1}$$

where $Z(\mathbf{X}) = \sum_{\mathbf{Y}} e^{\psi(\mathbf{Y},\mathbf{X})}$ is a normalization factor.

We utilize a conditional random field to estimate the most likely object silhouette within an image subwindow. The input, $\mathbf{X}$ is a graph (V, E), with one vertex corresponding to each pixel, and an edge between each adjacent pixel. $\mathbf{Y} = \{y_v, v \in V\}$ is the labeling of each pixel into one of two classes, object (1) or background (0). Additionally, we introduce a vector of unobservable index variables $\mathbf{W} = [w_f, w_b]$, related to the appearance of the foreground and background. Their role is described in the next section. The log potential function is then a sum of two terms, based on appearance and boundary.

$$\psi(\mathbf{Y}, \mathbf{W}, \mathbf{X}) = \psi^a(\mathbf{Y}, \mathbf{W}, \mathbf{X}) + \psi^b(\mathbf{Y}, \mathbf{X}) . \tag{2}$$

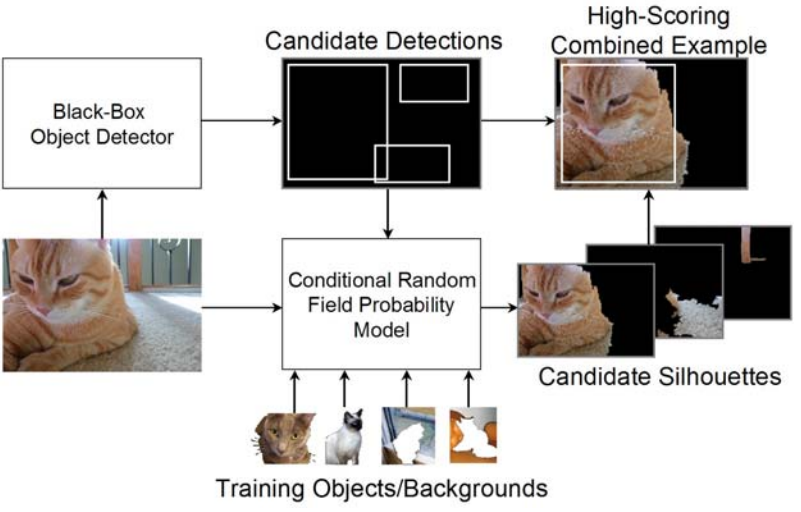

Fig. 2. Our approach

Appearance Potential. The appearance potential function must capture consistency between image pixel colors and the corresponding pixel label. It is important to note that object classes can be made up of a diverse set of individuals. Taking, for example, the variety of dog breeds, our probability model must assign high probability to labellings consistent with the coloring of either a German Shepherd or a Golden Retriever, but not when the foreground is a mixture of the two. For this reason we employ a nearest-neighbor like approach in our appearance potential function. We extract from the training images a set of object color histograms, one per individual, and introduce index variable w_f to represent the training individual to which the object labeled pixels are matched. Similarly, we introduce a second index, w_b, to represent the training background to which the background pixels match. Allowing w_f and w_b to vary separately allows us to correctly recognize cases where, for example, a Dalmatian is located outdoors, even if in training we only saw Dalmatians indoors and other dogs outdoors.

We define the appearance potential then as:

$$\psi^a(\mathbf{X}, \mathbf{W}, \mathbf{Y}) = \frac{1}{|V|} \sum_{v \in V} \left((y_v)\phi\left(col(v), hist_{f,w_f}\right) + (1 - y_v)\phi\left(col(v), hist_{b,w_b}\right)\right) . \quad (3)$$

$$\phi(c, hist) = \theta_1 log(hist(c)) . \quad (4)$$

where $hist(c)$ is the value of the corresponding color histogram at color c. $\phi(c, hist)$ can be interpreted as the negative description length of color c given the color probability distribution implied by $hist$.

Boundary Potential. Relative to the total number of edges in $\mathbf{X}$, we expect the number of edges joining foreground and background pixels to be few, and thus we define the boundary potential function to penalize large numbers of boundaries:

$$\psi^b(\mathbf{X}, \mathbf{Y}) = \sum_{e=(v_i,v_j)\in E} \frac{\theta_2 \,\mathrm{boundary}(y_i, y_j)}{\mathrm{width}(\mathbf{X}) + \mathrm{height}(\mathbf{X})} \,. \tag{5}$$

where $\mathrm{boundary}(y_i, y_j)$ = 1 if $y_i \neq y_j$ and 0 otherwise, width and height denote the width and height (in pixels) of the image subwindow. Here we normalize by their sum, as we expect the number of transition edges to be roughly proportional to this value.

CRF Construction. Given a candidate detection bounding box, we construct the input $\mathbf{X}$ for our conditional random field as follows. First we form an image subwindow by enlarging the bounding box by a factor of $\frac{1}{2}$ in each dimension in order to ensure that the entire object is contained should the bounding box be slightly too small. Because of the potentially large number of CRF's that are constructed and evaluated on a test image we employ the following approximations for efficiency. Instead of considering all possible (0,1) labelings over all pixels, we first apply an off-the-shelf segmentation algorithm [8] to segment the pixels into contiguous regions of similar color. Then, all pixels in each segment are constrained to have the same label. The segmentation algorithm parameters are dependent on the size of the bounding box ($(sigma, minsize)$ ranges from $(.5, 125)$ when bounding box < 16384 pixels to $(4, 1000)$ when bounding box > 65536 pixels). This strategy produces a roughly constant number of segments in the image subwindow, independent of its size (roughly 30 to 50). As we do not expect the object to extend into regions of the image outside of the CRF, all segments that extend beyond the image subwindow are constrained to be labeled as background.

Training. The individual object and background histograms are estimated from the silhouetted subset of training images from the PASCAL VOC 2007 dataset [3]. This is done simply by enlarging the annotated silhouette's bounding box as described above and recording each constituent pixel's color in the object or background color histogram. Colors are discretized into a 4096 bin smoothed histogram [9].

In order to learn CRF parameters θ, we construct a graphical structure and segmentation as described above for each object instance in the training images. Pixels are assigned foreground/background labels based on whether their associated segment is mostly in the silhouette or not. Thus we have a set $(\mathbf{X^t}, \mathbf{Y^t}), t = 1, ..., N$ of structured inputs and pixel-wise labelings. Indices w^t_{obj} and w^t_{back} are free parameters, but are constrained so as not to self-reference.

Ideally we could choose θ by maximizing the conditional log likelihood of the training data. Unfortunately, this requires calculating the normalizing factor $Z(\mathbf{X})$, which involves a sum over all pixel labelings, and so is not feasible. Thus, we employ an approximate method, contrastive divergence [10]. In contrastive divergence, a Markov chain is started at the correct labeling and run for a small number of steps (in our case one), and then the parameters are updated according to:

$$\theta^{s+1} = \theta^s + \lambda \left(\left\langle \frac{\partial \psi(\mathbf{X}, \mathbf{Y})}{\partial \theta} \right\rangle_{\mathbf{Y^0}} - \left\langle \frac{\partial \psi(\mathbf{X}, \mathbf{Y})}{\partial \theta} \right\rangle_{\mathbf{Y^1}} \right) \,. \tag{6}$$

where $\mathbf{Y_0}$ is the label distribution defined by the training data, and $\mathbf{Y_1}$ is the label distribution after one step in the Markov chain. This update has the effect of shaping the potential function locally to encourage the correct labeling. With the small number of parameters in our CRF model (two), we find that convergence is very fast (approximately 3000 iterations).

Application. For test images we proceed as follows. First the image is input to our black-box gradient based object detector, resulting in a set of candidate detections bounding boxes and associated scores, $\{(B_j, s_{B_j}), 1 \leq j \leq M\}$. This set is processed by a non-maximal suppression procedure, in which any candidate detections with stronger candidate detections nearby (overlapping by 80%) is removed. For each bounding box we construct a conditional random field by enlarging the bounding box and performing a size appropriate segmentation as described above.

The CRF defines a probability over all object/background labellings. Because of the large number of CRF's constructed (5 to 100 per image), evaluating each labeling is infeasible. Thus we find one high-probability labeling as follows. We initialize the labels such that all segments are assigned background except for those located entirely within the candidate bounding box B_j. w_f and w_b are chosen so as to maximize the color potential function. Then, segment labellings and indices are iterated through and changed as necessary so as to find a local maximum in the probability field. This set of $(\mathbf{X}, \mathbf{Y})$ pairs (one per candidate bounding box) is forwarded to the next stage for feature extraction.

3 High-Level Features

Color Consistency. Once the object silhouette, S, is estimated, we extract the foreground and background color histograms ($h_{S,f}$ and $h_{S,b}$). These histograms can then be compared to those from the training images. We generate two features (Equations 9 and 10) based on the consistency between these histograms and the matched training instances, and a third color feature to represent how salient the object/background distinction is (Equation 11):

$$f_{S,f,col} = -\left(\min_{w_f} D_{KL}\left(h_{S,f}, hist_{f,w_f})\right)\right) . \tag{7}$$

$$f_{S,b,col} = -\left(\min_{w_b} D_{KL}\left(h_{S,b}, hist_{b,w_b})\right)\right) . \tag{8}$$

where D_{KL} represents Kullback-Leibler divergence. Let $h_{S,all}$ be the histogram of all pixels in the candidate subwindow, and let $w_f = \arg\min_{w_f} D_{KL}(h_{S,f}, hist_{f,w_f}))$ and $w_b = \arg\min_{w_b} D_{KL}(h_{S,b}, hist_{b,w_b}))$. The saliency feature is then:

$$\begin{aligned} f_{S,colsal} = f_{S,objcol} + f_{S,backcol} \\ + D_{KL}\left(h_{S,all}, hist_{f,w_f})\right) + D_{KL}\left(h_{S,all}, hist_{b,w_b})\right) . \end{aligned} \tag{9}$$

Shape Consistency. We generate one feature $f_{S,shape}$ to represent the shape consistency between the estimated object silhouette and those seen in the training images. To do so, we first define a function, $\mathrm{ShapeRep}$, that inputs a silhouette and outputs a 500 dimensional vector of shape characteristic features. $\mathrm{ShapeRep}$ operates by placing a 10 x 10 grid (indexed by (i,j)) over the silhouette's bounding box, and calculating 1 shape element, $e_{i,j}^{shape}$, and 4 boundary elements, $e_{i,j,k}^{boundary}$ ($k = \{up, down, left, right\}$), per cell. $e_{i,j}^{shape}$ equals the fraction of foreground pixels in cell (i,j), and $e_{i,j,k}^{boundary}$ equals the fraction of total boundary edges located in cell (i,j) with orientation k. Finally, the vector is normalized to length 1. Again, we utilize a nearest neighbor approach in defining $f_{S,shape}$,

$$f_{S,shape} = \max_{T \in D_{sil}} \mathrm{ShapeRep}(S) \cdot \mathrm{ShapeRep}(T) \; . \tag{10}$$

where D_{sil} is the set of training silhouettes and their horizontal reflections.

Feature Combination. In order to combine the silhouettes/high-level features with the original set of bounding boxes/gradient scores, we define the following function:

$$\mathrm{match}(B_j, S_i) = \frac{\mathrm{IntersectionArea}(B_j, \mathrm{Box}(S_i))}{\mathrm{UnionArea}(B_j, \mathrm{Box}(S_i))} \; . \tag{11}$$

where $\mathrm{Box}(S_i)$ returns the minimal size box fully enclosing silhouette S_i. The match function ranges from 0 to 1 and measures the compatibility between a bounding box and silhouette. For each bounding box B_j we first identify those silhouettes S_i for which $\mathrm{match}(B_j, S_i)$ is greater than threshold α. Then, for each high level feature f_k, we set $f_{B_j,k}$ to the maximum value of f_k across all corresponding silhouettes.

Combined examples can then be assigned an overall confidence using a simple linear weighting of the bounding box score and high-level features of the combined example. In our empirical evaluations, detection methods are scored based on average precision over all recalls, and so we maximize this value over a set of withheld validation data (constraining all weights to be positive). Given the small number of features, we are able to implement this with a simple random walk through the weight space. Paramaeter α is optimized in this procedure as well.

4 Empirical Evaluation

We test our method on the set of animal categories from the PASCAL Visual Object Classes 2007 Detection Challenge dataset [3].[1] Animals (bird, cat, cow, dog, horse and sheep) are a difficult subclass of objects to detect because of their highly deformable nature. Object silhouettes are provided for roughly 5% of the images. As described in section 3.5, training image silhouettes are used to acquire object/background color

[1] We choose the 2007 dataset as the PASCAL VOC 2008 test data is not publicly available at the time of writing.

histograms and shape representations. This procedure results in approximately 50 training examples per class.

For the black-box gradient-based approach, we use the latest object detection system of Felzenszwalb et al., [1,11]. This system represents the state of the art earning the highest published overall detection rates for both the VOC 2007 and VOC 2008 detection challenges. We take advantage of their publicly available detection source code. While their model training code is unavailable at the time of writing, they made available models for each of the 20 VOC 2007 object classes. As the gradient based system has already been trained on all training images, in order to learn the linear combination parameters we randomly select half of the test images for validation. The parameters are chosen so as to maximize the average precision across the validation set. Finally, the combined approach is evaluated against the remaining test images. We repeat this process 6 times and take the average results.

In the VOC detection challenge, a correct detection is defined as one for which there exists an annotated ground truth bounding box of the same class such that $\frac{Areaof Intersection}{Areaof Union} > \frac{1}{2}$, however, only one detection is permitted for each ground-truth bounding box, all subsequent detections are considered false positives. We compare the performance of our combined approach (Low-level + High-Level Features) to that of the gradient based detector alone (Low-Level Features). For each system, the candidate detections are ranked based on confidence, non-maximal suppression is performed, and the average precision over all recalls is evaluated. Results appear in table 1.

Table 1. Our Detection Rates vs. [11](Average Precision). †: Best in VOC 2007 Detection Challenge (Cow, Horse) and VOC 2008 Detection Challenge (Bird).

Class	State of the Art (Low-Level) [11]†	Our System (Low-Level + High-Level)	Change
Bird	.0193	.0242	25.4%
Cat	.115	.120	4.3%
Cow	.148	.144	-2.5%
Dog	.098	.117	19.4%
Horse	.362	.372	2.9%
Sheep	.245	.248	1.4%

We see that the inclusion of high level features improves the results substantially for the two most difficult of the six classes (bird and dog). Potentially, the wide variety of perspective and poses for these objects renders the gradient based detector volatile, and the global color consistency features serve as an important cue. Improvements in other classes are modest (cat, horse, sheep), and the high-level features only decreases performance in one case (cow), and by a small amount. This is likely due to the low dimensional nature of the high-level features, which help the overall system resist overfitting the validation data. Example silhouettes are shown in Figure 3.

Fig. 3. Example estimated object silhouettes. The last three columns illustrate the matched training instance for each of object color, background color, and object shape.

5 Conclusion

We present an approach to object detection that enhances the low-level local feature approach popular in today's literature with a high-level feature extraction stage to

increase accuracy in object detection. In our approach candidate detections are fed into a conditional random field probability model to identify the most likely silhouette corresponding to each detection. Based on these silhouettes, color and shape consistency features are extracted and combined in a simple linear weighting for enhanced detection accuracy. We demonstrate the merit of our approach on the difficult animal classes from the PASCAL VOC 2007 object detection dataset.

References

1. Felzenszwalb, P., McAllester, D., Ramanan, D.: Discriminatively trained, multiscale, deformable part models. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
2. Zhang, H., Berg, A., Maire, M., Malik, J.: Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2 (2006)
3. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC 2007) Results (2007), `http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html` (accessed November 1, 2008)
4. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2008 (VOC 2008) Results (2008), `http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html` (accessed November 1, 2008)
5. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning (2001)
6. Quattoni, A., Collins, M., Darrell, T.: Conditional random fields for object recognition. In: Neural Information Processing Systems (2004)
7. He, S., Zemel, R., M., C.P.: Multiscale conditional random fields for image labeling. In: IEEE Conference of Computer Vision and Pattern Recognition (2004)
8. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. International Journal on Computer Vision 59(2) (2004)
9. Forsyth, D., Ponce, J.: Computer Vision: A Modern Approach. Prentice Hall, Englewood Cliffs (2003)
10. Hinton, G.: Training products of experts by minimizing contrastive divergence. Neural Comp. 14, 1771–1800 (2002)
11. Felzenszwalb, P., McAllester, D., Ramanan, D.: Discriminatively trained mixtures of deformable part models. In: PASCAL Visual Object Challenge (2008)