Location Constrained Pixel Classifiers for Image Parsing with Regular Spatial Layout

Kang Dang
kangdang@gmail.com
Junsong Yuan
jsyuan@ntu.edu.sg

School of Electrical and Electronic
Engineering, Nanyang Technological
University, Singapore 639798

Location is useful for a variety of image parsing problems with regular spatial layout, such as pedestrian parsing after detection, street view scene parsing and medical image segmentation. This paper proposes a novel way to leverage both location and appearance information for pixel labeling (Fig. 1).

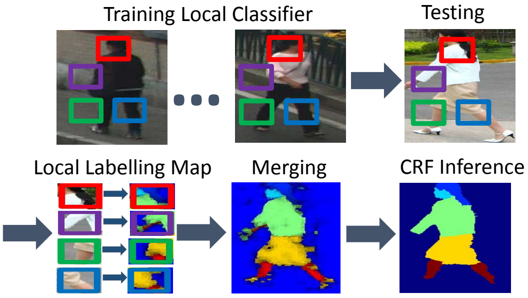

Figure 1: Overview of our method. (1) At each location we train a position dependent local pixel classifier with training pixel samples from its neighborhood region represented by a patch. (2) Assume the training and testing images have similar layout, the trained classifier is used for the same region in the testing images. (3) To ensure smoothed labeling results, we allow the local classifiers to overlap with each other, such that each pixel will be voted by multiple local classifiers. The final score is the average score of all engaged local classifiers. (4) The final result is obtained after a proper discretization of the labeling map with a conditional random field (CRF).

Existing approaches solve the pixel labeling problem with a single global model. In other words, they learn a single global pixel classifier for the entire image space, and all pixels of the image are used to train the classifier. In contrast, at each image location we learn a location constrained classifier, i.e. local classifier. Since each local pixel classifier is learned by only using the pixels in a local neighborhood, it is expected to better fit the local pixel distribution and capture local discriminative information. To prevent local classifiers overly depending on the image location and to improve the generalization, the neighborhood scale of local learning is important. We justify the significance of the neighborhood scale via the following theoretical studies.

Probabilistic Analysis. Given a pixel's central position (x,y) and its associated feature vector $\mathbf{f}$, our goal is to predict the class label L of that pixel. We are interested in learning a number of local classifiers $p_{\mathcal{N}}(L \mid \mathbf{f})$ at different spatial locations. $\mathcal{N}(x,y,s)$ stands for a local image neighborhood, which is a patch centered at (x,y) and of width $s \times \mathcal{W}$ and height $s \times \mathcal{H}$, where s is the neighborhood scale and $\mathcal{W}$ and $\mathcal{H}$ is the width and height of the image. In other words, the training set for each local classifier is $\{(L_i, \mathbf{f}_i) \mid \forall (x_i, y_i) \in \mathcal{N}(x,y,s)\}$. We show the local classifier approximates the following conditional distribution:

$$p_{\mathcal{N}(x,y,s)}(L \mid \mathbf{f}) \propto \sum_{(x,y)\in\mathcal{N}(x,y,s)} p(L \mid \mathbf{f},x,y)\, p(\mathbf{f} \mid x,y). \qquad (1)$$

We see the proposed local classifier $p_{\mathcal{N}}(L \mid \mathbf{f})$ is a spatially smoothed version of the global classifier $p(L \mid \mathbf{f},x,y)$ in a local neighborhood, where the weight $p(\mathbf{f} \mid x,y)$ characterizes the dependency of the observed feature $\mathbf{f}$ at the pixel location (x,y). The neighborhood scale s plays an important role in building the local classifier. On one hand, when the local neighborhood contains only a single pixel, *i.e.*, $s = 0$, our local classifier degenerates into: $p_{\mathcal{N}(x,y,0)}(L \mid \mathbf{f}) = p(L \mid \mathbf{f},x,y)$. On the other hand, when the local neighborhood expands to the entire image, *i.e.*, $s = 1$, it becomes $p_{\mathcal{N}(x,y,1)}(L \mid \mathbf{f}) = p(L \mid \mathbf{f})$, which indicates position information (x,y) is not utilized at all. Our proposed classifier is a compromise between these two ends.

Bias-Variance Trade-Off. We discuss the implication of choosing an appropriate neighborhood scale s from the perspective of bias-variance analysis. Our main conclusion is a theorem stating that under certain assumptions, testing error variance monotonically decreases with the neighborhood scale s. In addition, our simulation shows that the bias increases with the neighborhood scale. Thus, an appropriate neighborhood scale is essential for balancing the bias and variance and minimizing the testing error.

	Penn-Fudan	PPSS
Feature Only	45.1	31.8
$(\mathbf{f},x,y)$ + SVM	54.3	39.7
$(\mathbf{f},x,y)$ + Boosting	60.3	45.1
Product of Expert	52.6	45.1
Ours	**63.1**	**53.5**

Penn-Fudan	
SBP[1]	57.3
P&S[4]	55.0
DL[3]	59.9
Ours	**63.1**
PPSS	
DDN[3]	47.2
Ours	**53.5**

Table 1: Benchmark results for Penn-Fudan and PPSS dataset. The performance metric is the average intersection over union(IOU) score over all labels. We compare our approach with three common methods of feature fusion and the state of arts. (1) $(\mathbf{f},x,y)$ + SVM: we concatenate feature and position information together to form $(\mathbf{f},x,y)$, and put it into a SVM classifier. (2) $(\mathbf{f},x,y)$ + Boosting: we put the concatenated feature vector $(\mathbf{f},x,y)$ into a joint boosting classifier. (3) Product of Experts: the merge is done by multiplying the two posterior probability map with weighting: $\frac{p(L|x,y)^k p(L|\mathbf{f})^{(1-k)}}{Z}$, where k is between 0 and 1, and Z is a normalization constant.

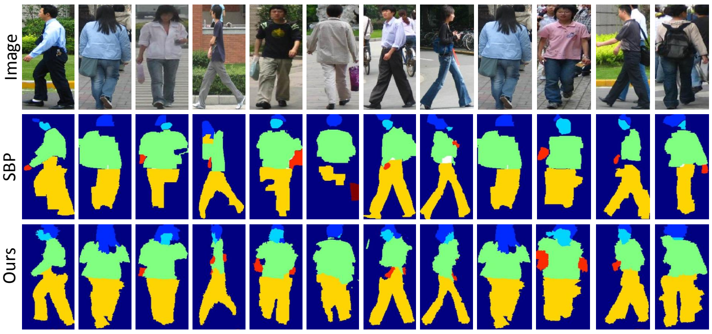

Figure 2: Image results from Penn-Fudan dataset. Visual quality is generally better than SBP[1].

Experiments. Our experimental evaluation is performed on two pedestrian parsing datasets Penn-Fudan [1] and PPSS dataset [3] as well as Weizmann horse segmentation[2]. Albeit simple, our proposed local learning works surprisingly well in these challenging image parsing problems. Some quantitative and qualitative results for pedestrian parsing datasets are shown in Table. 1 and Fig. 2. It confirms the advantages of our local classifiers which are better adapted to the local image characteristics than a global classifier.

[1] Yihang Bo and Charless C Fowlkes. Shape-based pedestrian parsing. In *CVPR*. IEEE, 2011.

[2] Eran Borenstein and Shimon Ullman. Class-specific, top-down segmentation. In *ECCV*. Springer, 2002.

[3] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian parsing via deep decompositional network. In *ICCV*. IEEE, 2013.

[4] Ingmar Rauschert and Robert T Collins. A generative model for simultaneous estimation of human body shape and pixel-level segmentation. In *ECCV*. Springer, 2012.