Image Parsing with a Wide Range of Classes and Scene-Level Context

Marian George
Department of Computer Science, ETH Zurich, Switzerland

Scene parsing is the assignment of semantic labels to each pixel in a scene image. The recognition rate of parsing methods significantly varies among different types of classes. Background classes are usually recognised with a high rate (e.g., road and building). Foreground classes (e.g., person and sign) represent salient image regions, but are frequently misclassified. Recently, nonparametric image parsing methods have been proposed [1, 2, 3, 4, 6] to handle the increasing number of scene categories and semantic labels. First, an image retrieval set is extracted, which contains the training images that are most visually similar to the query image. The number of candidate labels for a query image is restricted to those in the retrieval set only. Second, classification likelihood scores of superpixels are obtained through visual features matching. Finally, context is enforced through minimizing an energy function which combines the data cost and knowledge about the classes co-occurences. Image retrieval is regarded as a very critical step [3, 6]; if the true labels are not included in the retrieved images, there is no chance to recover from this error later in the pipeline.

We propose a novel nonparametric image parsing method that achieves better overall accuracy with better coverage of rare classes. Our contributions: (1) Improving the likelihood scores of labels at superpixels through combining classifiers. Our system combines the output probabilities of multiple classification models to produce a more balanced score for each label at each superpixel. (2) Incorporating semantic context in a probabilistic framework. To avoid the elimination of relevant labels in the filtering step, we do not construct a retrieval set. Instead, we use label costs learned from the contextual correlation of labels in similar scenes to achieve better results.

Fusing Classifiers Our method is inspired from ensemble classifier methods that combine multiple classifiers to reach a better decision. Such techniques are specifically useful if the classifiers are different, i.e., the error reduction is related to the uncorrelation between the trained models [5]. To this end, we train 4 Boosted Decision Tree (BDT) models with the following training data criteria: (1) an unbalanced subsample of all classes, (2) a balanced subsample of all classes, (3) a balanced subsample of classes occupying an average of less than $x\%$ of their images, and (4) a balanced subsample of classes occupying an average of less than $\lceil x/2 \rceil$ % of their images. The motivation is to reduce the correlation between the trained models. While the unbalanced classifier mainly misclassifies the foreground classes, the balanced classifiers recover some of these classes while making more mistakes on the background ones. By combining the likelihoods from all the classifiers, a better decision is reached that covers more classes.

The final cost of assigning a label c to a superpixel s_i can then be represented as the combination of the likelihood scores of all classifiers:

$$D(l_{s_i} = c|s_i) = 1 - \frac{1}{1+e^{-L_{comb}(s_i,c)}} \quad (1)$$

where $L_{comb}(s_i,c)$ is the combined likelihood score obtained by the weighted sum of the scores from all classifiers:

$$L_{comb}(s_i,c) = \sum_{j=1,2,3,4} w_j(c) L_j(s_i,c), \quad (2)$$

where $L_j(s_i,c)$ is the score from the j^{th} classifier, and $w_j(c)$ is the normalized weight of the likelihood score of class c in the j^{th} classifier.

We learn the weights $\mathbf{w} \equiv [w_j(c)]$ of all classes C in offline settings using the training set. The weight $\dot{w}_j(c)$ of class c for the j^{th} classifier is computed as the average ratio of the sum of all likelihoods of class c, to the sum of all likelihoods of all classes $c_i \in C \backslash c$ of all superpixels $s_i \in S$. The normalized weight $w_j(c)$ of class c can then be computed as: $w_j(c) = \dot{w}_j(c) / \sum_{j=1,2,3,4} \dot{w}_j(c)$. Normalizing the output likelihoods in this manner gives a better chance for all classifiers to be considered in the result.

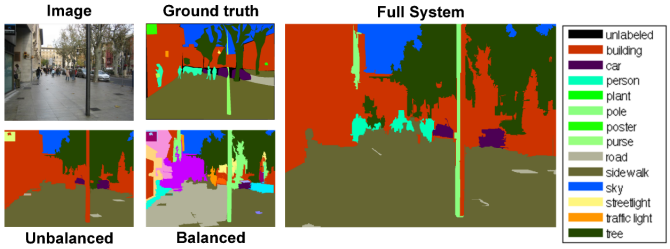

Figure 1: Image parsing by combining likelihoods from unbalanced and balanced classifiers to cover a wider range of classes.

Scene-Level Context We incorporate semantic context through using label statistics instead of global visual features. The intuition is that ranking by visual features often fails to retrieve similar images on the scene level [4, 6]. However, ranking by label statistics, given a relatively good initial labeling, retrieves more semantically similar images to remove outlier labels and recover missing labels in a scene. Our approach does not limit the number of labels to those present in the retrieval set but instead uses the set to compute the likelihoods in a k-nn fashion. The likelihoods are normalized and smoothed to give a chance to labels not in the retrieval set.

For a given test image, let $T \subset C$ be the set of unique labels which appear in the initial labeling L. We model the conditional distribution $P(c|T)$ over class labeling C given T. We compute $P(c|T) \forall c \in C$ in a K-nn fashion:

$$P(c|T) = \frac{(1+n(c,K_T))/n(c,S)}{(1+n(\bar{c},K_T))/|S|}, \quad (3)$$

where K_T is the K-neighborhood of T, $n(c,X)$ is the number of superpixels with label c in X, $\bar{c}$ is all labels except c, and $|S|$ is the total number of superpixels in the training set. We add a smoothing constant of value 1.

To get the neighborhood K_T, we rank the training images by their distance to the query image. The distance between two images is computed as the *weighted* size of intersection of their class labels, intuitively reflecting that the neighbors of T are images with many shared labels with those in T. We assign a different weight to each class in T in such a way to favor less-represented classes. Once we obtained the likelihoods $P(c|T)$, we can define a label cost $H(c) = -log(P(c|T))$. Our final energy function becomes:

$$E(L) = \sum_{s_i \in S} D(l_{s_i} = c|s_i) + \lambda \sum_{(i,j) \in A} V(l_{s_i}, l_{s_j}) + \sum_{c \in C} H(c).\delta(c), \quad (4)$$

where $\delta(c)$ is the indicator function of label c, $D(l_{s_i} = c|s_i)$ is the data cost, and $V(l_{s_i}, l_{s_j})$ is the smoothing cost.

Our system achieves state-of-the-art per-pixel recognition rates on two large-scale datasets. We achieve 81.7% per-pixel accuracy and 50.1% per-class accuracy on SIFTflow [2]. Our fusing classifiers step boosts the per-class accuracy by 15% over the baseline. On LMSun [4], we achieve 61.2% per-pixel accuracy and 16% per-class accuracy.

[1] D. Eigen and R. Fergus. Nonparametric image parsing using adaptive neighbor sets. In *ECCV*, 2008.

[2] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 33(12):2368–2382, 2011.

[3] G. Singh and J. Košecká. Nonparametric scene parsing with adaptive feature relevance and semantic context. In *CVPR*, 2013.

[4] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. *IJCV*, 101(2):329–349, 2013.

[5] K. Tumer and J. Ghosh. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(3-4):385–404, 1996.

[6] J. Yang, B. Price, S. Cohen, and M.-H. Yang. Context driven scene parsing with attention to rare classes. In *CVPR*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.