Semantic Image Classification Using Consistent Regions and Individual Context

Stefan Kluckner
kluckner@icg.tugraz.at
Thomas Mauthner
mauthner@icg.tugraz.at
Peter M. Roth
pmroth@icg.tugraz.at
Horst Bischof
bischof@icg.tugraz.at

Institute for Computer Graphics and Vision
Graz University of Technology
Austria

This paper proposes an efficient approach for semantic image classification by integrating additional contextual constraints such as class co-occurrences into a randomized forest (RF) classification framework. The RF classifier performs an initial yet local classification on the pixel level by using powerful covariance matrix based descriptors as feature representation. Furthermore, we exploit multiple unsupervised image partitions to provide a reliable spatial region support and to capture the real object boundaries. An information theoretic driven approach detects consistently classified regions and generates a representative segmentation incorporating the classification result on the pixel level. Moreover, we use a conditional random field formulation to obtain a final labeling including context information individually generated for each test image. To illustrate state-of-the-art performance, we run experiments on the two versions of the MSRC dataset with 9 and 21 object classes and on the PASCAL VOC2007 image collection.

Covariance Region Descriptors. We use powerful yet compact covariance regions descriptors [6] as the feature representation within a RF classifier by applying a simple matrix vectorization similar to [1]. This representation then directly integrates arbitrary feature cues, such as color and filter responses.

Consistently Classified Regions. Since our local RF classification strategy yields a class distribution at each pixel independently, we aim to group the obtained information according to its spatial relationship. Following the concepts presented in [4, 5], multiple segmentations are generated to provide a huge pool S of probable connected pixels. For each segmented region $s_i \in S$, we group the individual pixel classifications yielding a final region class distribution. In order to select consistently classified regions, we compute the Shannon entropy over the summarized distribution. Given a computed class distribution $P(\mathbf{c}|s_i)$ of a segmented region s_i, we define a consistency measurement based on an entropy computation according to

$$H(s_i) = -\sum_{j=1}^{|c|} p_j \log p_j, \qquad p_j = \frac{\log P(c=j|s_i)}{\sum_{k=1}^{|c|} \log P(c=k|s_i)}, \tag{1}$$

where $|c|$ is the number of object classes. Considering the set of generated segmentations S and the corresponding consistency measurements $H(S)$, the final image partition is constructed as follows: For each pixel q in the image I we estimate the segmentation index $i_q^* = \arg\min_{q \in s_i} H(s_i)$ by minimizing the obtained entropies over all segments in S that include q. These indexes are stored in an image structure and provide the final partition for the CRF stage, that incorporates the contextual knowledge.

Individual Context. As a last step, we exploit the structure of the RF to generate individual context information for each image. Following the idea of Gall [2] we store additional information, such as a probable class occurrence configuration, in the tree's leaf nodes. Each feature instance is extended to include the classes occurring in the current training images considering the ground truth labeling. At runtime the classifier is evaluated at the pixel level by parsing down the extracted feature representation in the forest. The learned symmetric co-occurrence matrix in the leaf node votes for an overall possible class configuration, which is directly applied to the CRF stage as semantic contextual knowledge. In this work we apply the efficient primal-dual strategy of [3] to minimize the energy within a region adjacency graph.

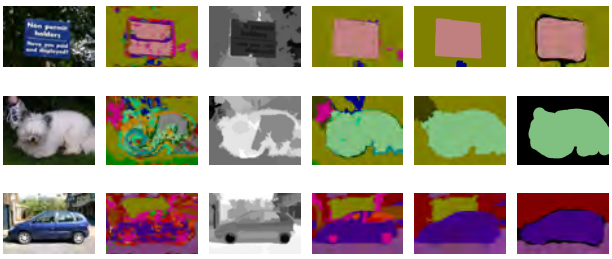

Figure 1: Visual results selected from the classification procedure on the MSRC dataset. From left to right: The color images, the initial RF classifications, the computed entropies minimized over all segmentations, the semantic labeling results using multiple segmentations, the CRF cleaned final classifications, and the ground truth.

Experiments. We evaluate the processing stages of our approach on the pixel level and compare the results to state-of-the-art performance. We show initial results obtained by the RF classifier, the classification rates by grouping the pixel using multiple segmentations and the rates by using a CRF formulation for the integration of contextual constraints thus allowing to asses the importance of the different steps. A quantitative evaluation shows that our feature representation, that directly integrates several cues, results in a reliable initial classification. The incorporation of unsupervised multiple segmentations significantly improves the accuracy. In addition, the final integration of individually generated context information yields competitive results on the MSRC and VOC2007 datasets. Figure 1 shows some visual results obtained by our approach.

Acknowledgments. This work was supported by the Austrian Science Fund Projects P18600 and W1209 under the doctoral program Confluence of Vision and Graphics, by the FFG projects APAFA (813397) and AUTOVISTA (813395), financed by the Austrian Research Promotion Agency, and by the Austrian Joint Research Project Cognitive Vision under the projects S9103-N04 and S9104-N04.

[1] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007.

[2] Juergen Gall and Victor Lempitsky. Class-specific hough forests for object detection. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, 2009.

[3] Nikos Komodakis and Georgios Tziritas. Approximate labeling via graph cuts based on linear programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1436–1453, 2007.

[4] Tomasz Malisiewicz and Alexei A. Efros. Improving spatial support for objects via multiple segmentations. In *Proceedings British Machine Vision Conference*, 2007.

[5] Caroline Pantofaru, Cordelia Schmid, and Martial Hebert. Object recognition by integrating multiple image segmentations. In *Proceedings European Conference on Computer Vision*, 2008.

[6] Fatih Porikli, Oncel Tuzel, and Peter Meer. Covariance tracking using model update based on lie algebra. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, 2006.