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

Abstract

This paper proposes an efficient approach for semantic image classification by integrating additional contextual constraints such as class co-occurrences into a randomized forest classification framework. The randomized forest 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 [21] dataset with 9 and 21 object classes and on the PASCAL VOC2007 [6] image collection.

1 Introduction

The problem of semantic description is still a largely unsolved task since there are huge visual variabilities of natural objects in available images. Typically, illumination, viewpoint and scale changes, and occlusions complicate the problem of finding a meaningful object representation within a classification process. Thus, recently the topic of semantic classification and segmentation is of major scientific interest in the computer vision community.

Local strategies, using supervision, aim to describe every pixel within a small neighborhood of the image space [12, 20, 21, 22, 24, 25]. Once a meaningful explanation of a pixel/region level is found, contextual constraints are integrated to find a final classification. These contextual constraints capture the probability of class occurrences within an image and also provide a smooth labeling in a spatial neighborhood. Conditional Markov random field (CRF) formulations [11] are widely adopted to include these contextual constraints [16,

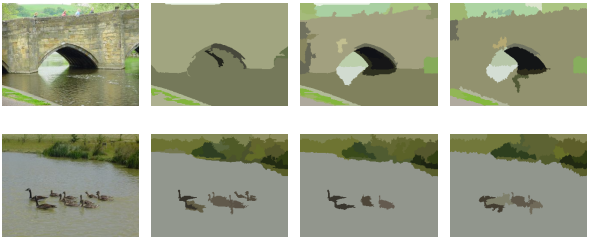

Figure 1: Examples of image partitions obtained by varying the parameter settings of a mean-shift segmentation [3]. The resulting image partitions do not guarantee perfect object boundaries. In some cases the objects such as the birds are not segmented or the bridge is mixed up with water regions.

[18, 20, 21, 24]. Due to missing perfect single image partitions [23], there is a trend to perform object classification by integrating multiple image segmentations [8, 10, 14, 16, 19]. In Figure 1, samples of different image segmentations are shown. Obviously, the varying configurations for the segmentation procedure capture the real object boundaries only to some extent,*e.g.*, the bridge is fused with water or tiny birds are missing in some partitions.

In [14], Malisiewics and Efros investigated the application of multiple segmentations to provide spatial support. They showed that a correct spatial support improves the recognition performance significantly. Pontafaru *et al.* [16] integrated multiple segmentations to obtain a final image partition that captures the real object boundaries. A classifier is trained on features that are extracted from segmented regions. Kohli *et al.* [10] introduced multiple segmentations as additional potentials within a CRF framework to enforce consistently labeled regions. Galleguillos *et al.* [8] used multiple instance learning, where the classifier is trained on a bag-of-word model, integrating computed stable segmentations.
Randomized forests (RF) [2] are well suited for multi-class object recognition due to accuracy and robustness to label noise [15, 20, 22]. Shotton *et al.* [22] proposed an RF framework incorporating simple semantic textons for initial classification. Schroff *et al.* [20] extended this approach by exploiting several feature cues and by applying more sophisticated split criteria within the forest. Furthermore, the structure of the trees simply enables to extract additional information such as hierarchical histograms [15, 22] or spatial features for object detection [7].

This paper has three main contributions: First, we use powerful yet compact covariance regions descriptors [17] as feature representation within an RF classifier by applying a simple matrix vectorization. This representation then directly integrates arbitrary feature cues such as color and filter responses. Second, we investigate how multiple segmentations, provided by [3, 6], can be integrated using the raw classification on the pixel level. The main idea is based on identifying regions that provide a consistent classification (we use an entropy measure). Minimizing the entropy over all different segmentations yields a final image partition that is represented by a region adjacency graph. Third, we exploit the structure of the RF to generate individual context information for each image. Following the idea of Gall *et al.* [7] we store additional information such as a probable class occurrence configuration, in the leaf nodes. For each test image the generated context information, represented by a co-occurrence matrix [18], is integrated by using an efficient CRF formulation [11] to obtain the final labeling of the regions adjacency graph.

The remaining part of the paper is structured as follows: First, in Section 2 we highlight the general classification pipeline that provides the class distribution on the pixel level. Additionally, the section illustrates the idea of extracting additional context information. Section 3 describes the generation of multiple image segmentations and how a final partition using the entropy minimization can be obtained. In Section 4 the integration of the context information is discussed. Section 5 presents the experimental evaluation on standard datasets. Finally, Section 6 concludes the paper and gives an outlook on future work.

2 Initial Classification on Pixel-Level

The first step of our approach involves a semantic classification procedure that provides an accurate explanation for each pixel in the image. Contrary to [20], where multiple feature channels are combined in a computationally costly manner, we directly integrate various cues in a single, compact feature representation based on covariance descriptors. In the following, we first highlight the idea of this feature representation and then we illustrate the application to the RF framework. In addition, we show how to exploit the structure of the RF for generating individual contextual constraints per image.

2.1 Covariance Features

To obtain a reliable local classification on the pixel level, we apply a strong feature representation based on covariance matrix descriptors. Covariance descriptors [17] are now widely used in object detection and tracking, since they can be efficiently computed and allow to capture the variance feature channels and correlation between them. Given a color image with three layers $I \in \mathbb{R}^{w\times h\times 3}$, we can generate a d-dimensional feature image $F \in \mathbb{R}^{w\times h\times d}$ from I using a mapping function $F(q) = \gamma(I,q)$, where $q = (x,y)$ defines an arbitrary pixel coordinate. Then, any rectangular $n \times n$ dimensional region $R \subseteq F$ can be represented by a $d \times d$ covariance matrix

$$\mathbf{C}_R = \frac{1}{|R|-1}\sum_{q\in R}\left(F(q)-\mu\right)\left(F(q)-\mu\right)^T, \tag{1}$$

where $\mu \in \mathbb{R}^{1\times d}$ is the sample mean vector. In this work, the mapping function $\gamma(I,q)$ provides the CIELab color values, the absolute values of the first and second order derivatives, the angle, and the magnitude of the gradients. The resulting feature vector $f \in \mathbb{R}^{1\times 9}$ for each pixel is defined as $f = \left[\begin{array}{cccccccc} L & a & b & |I_x| & |I_y| & |I_{xx}| & |I_{yy}| & \arctan\left(\frac{I_y}{I_x}\right) & \sqrt{I_x^2+I_y^2}\end{array}\right]$. The derivatives are computed on the L-color channel. Since the space of covariance matrices is non-Euclidean, these descriptors can not be directly used in RFs.

To overcome this problem, in [17] iterative numerical procedures using an affine-invariant Riemannian metric are applied. In contrast, in this paper we aim to use a more efficient computation scheme based on a log-Euclidean Riemannian metric [1] exploiting the underlying structure of symmetric positive definite (SPD) matrices. Covariance matrices $\mathbf{C}_R$ are positive semi-definite by definition. Hence, by simple regularizing $\mathbf{C}_R = \mathbf{C}_R + \varepsilon\mathbf{I}_d$, where $\mathbf{I}_d$ is the d-dimensional identity matrix and $\varepsilon =$ 1e-9, $\mathbf{C}_R$ becomes symmetric positive definite. In fact, the space of positive definite matrices $Sym^+(d)$ lies on a connected Riemannian manifold describing a Lie group. However, under the log-Euclidean Riemannian metric the Lie group structure of SPD matrices can be extended to a Lie Algebra, which also defines a vector

space, that can be described by a Euclidean metric. More details can be found in [1]. Thus, given a region covariance matrix $\mathbf{C}_R \in Sym^+(d)$, we apply the log-Euclidean mapping:

$$\log(\mathbf{C}_R) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} (\mathbf{C}_R - \mathbf{I}_d)^k \quad = \quad \mathbf{U} \log(\mathbf{D}) \mathbf{U}^T, \tag{2}$$

where $\mathbf{C}_R = \mathbf{U}\mathbf{D}\mathbf{U}^T$ is the eigenvalue decomposition and $\log(\mathbf{D})$ is the diagonal matrix of the eigenvalue logarithms. Since $\log(\mathbf{C}_R)$ has a vector space structure under the log-Euclidean metric, we can unfold $\log(\mathbf{C}_R)$ into a feature vector without loosing any information about $\mathbf{C}_R$. Therefore, we can represent the information given by a $d \times d$ dimensional covariance matrix $\mathbf{C}_R$ by a vector $\mathbf{v} \in \mathbb{R}^{1 \times d^2}$. Following [13], where this vectorized form of log-Euclidean covariance matrices has been used for learning incremental subspaces in tracking, we apply the feature representation $\mathbf{v}$ to our RF classifier.

2.2 Randomized Forest Classification

RFs [2] provide an efficient technique to obtain an averaged class distribution and can also be used to extract hierarchical data for further improvement of the classification [15, 22]. A forest consists of an ensemble of T binary decision trees, where the nodes of each tree include split criteria that give the direction of branching left and right down the tree until a leaf node is reached at a given maximum depth D. A leaf node l_i then contains the likelihood $P(\mathbf{c}|l_i)$ given by the labels of the visible training examples. An averaging over all decisions in the forest yields the resulting accumulated class distribution obtained at each evaluated pixel location according to $P(\mathbf{c}|L) = \frac{1}{T}\sum_{i=1}^{T} P(\mathbf{c}|l_i)$. Given a training set $\mathbf{V} = \{\mathbf{v}^1, \ldots, \mathbf{v}^N\}$, each tree is learned on a randomly extracted subset $\mathbf{V}' \subset \mathbf{V}$. Additionally, the target labels are extracted by considering the ground truth data and are directly assigned to the feature representation. A concrete training sample $(\mathbf{v}^i, c_i)$ consists of the feature $\mathbf{v}^i$ and the assigned target label c_i. The learning proceeds from the root node top-down by splitting the available subset at each node into tiled left and right feature sets. The split criteria in the non-leaf nodes minimize the sample weighted information gain ratio [22] of the class distribution in currently available subsets of the training data. We follow a similar strategy as proposed in [2, 22] to perform the splitting decisions by comparing two randomly chosen attributes v^i_j and v^i_k of the available feature sample $\mathbf{v}^i \in \mathbf{V}'$:

$$v^i_j - v^i_k = \begin{cases} > 0, & left\ branch \\ \leq 0, & right\ branch \end{cases}. \tag{3}$$

Here, j and k denote the indexes of the selected attributes that maximize the information gain considering the training labels. We take the same numbers of split node tests as suggested in [22].

2.3 Exploiting the Tree Nodes

Recently, Gall and Lempitsky [7] showed that the structure of RFs can be exploited to store additional information like coordinate offsets for the task of object detection. We propose a similar mechanism to generate a voting for probable object class occurrences individually for each test image. To obtain the integration of an image specific probable class configuration, each feature instance is first extended to include the classes occurring in the current

training images considering the ground truth labeling. Therefore, we construct a binary vector that masks a possible occurrence of related classes and assign it to the resulting feature vector. While recursively learning the trees, taking into account the feature representation (see Section 2.1) and the split criteria defined in Equation 3, the feature instances end up in a leaf node where the assigned binary information is accumulated for each permutation of possible class occurrences. Once the forest is trained, the classifier is evaluated at the pixel level by parsing down the extracted feature representation $\mathbf{v}^i$ in the forest and summarizing the class distribution to obtain an averaged likelihood at each pixel location. Additionally, the learned symmetric co-occurrence matrix in the leaf node votes for an overall possible class configuration $\theta \in \mathbb{R}^{|c|\times|c|}$, which is directly applied to the CRF as semantic contextual knowledge. Here, $|c|$ is the number of object classes. Section 4 highlights the details for the integration of the co-occurrence information into the classification pipeline.

3 From Pixel-Level to Region Classification

Since our local RF classification strategy yields a class distribution for each pixel in the image independently, we aim to group the obtained information according to its spatial relationship. Following the concepts presented in [14, 16], multiple segmentations are generated to provide a huge pool of probable connected pixels. For each segmented region, 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. Taking into account the minimum entropy over all segmentations for each pixel, a final partition is obtained by assigning the index of the corresponding region.

3.1 Generate Multiple Segmentations

As shown in [8, 14, 16] an application of different approaches for unsupervised image segmentation captures the huge variety in color, scales, texture, etc. In this work we employ two segmentation approaches (*i.e.* the graph-based method proposed by Felzenszwalb and Huttenlocher [6] and the mean-shift approach [3]), which are selected due to public availability, efficiency, popularity and the use of different techniques for image partitioning. Please note that any other segmentation technique could be employed, since we are only interested in generating probable image partitions. For each test image, multiple segmentations are produced using these methods with varying parameter settings. In our implementation we consider an overall number of 15 segmentations. The first six segmentations are generated using mean-shift by setting the spatial band to $b_s = \{5,9\}$ and the range band to $b_r = \{6,12,18\}$. Using the graph-based segmentation, we get the remaining nine partitions by varying $\sigma = \{0.5,1.0,1.5\}$ and $k = \{100,300,600\}$. The minimum region size is set to 200 pixels for both segmentation methods. An obtained segment $s_i = \{q_1,\dots q_K\}$ can be seen as a list of pixel coordinates that describes the region i. In the following sections, we denote the obtained pool of segmented regions as $S = \{s_1,\dots s_N\}$, where N defines the resulting number of regions produced by the 15 different segmentation procedures for a given test image I.

3.2 Grouping the Pixel-Level Information

Instead of classifying all regions in the set S, *e.g.*, using an SVM [16], we aim to generate an overall class distribution by summarizing the local results on the pixel level within a given region $s_i \in S$. Taking into account the pixel level classification, we obtain an overall region class distribution in terms of log likelihoods with

$$\log P(\mathbf{c}|s_i) = \sum_{q \in s_i} \log P(\mathbf{c}|L,q)\,. \tag{4}$$

In contrast to [16], where the individual region class labels of different segmentations vote in an accumulation procedure, we determine the classification consistency for each region and construct the final segmentation providing the input for our context integration process. Given the class distribution in terms of *log* likelihoods for each region $s_i \in S$, we first introduce a normalization step to enforce $\sum_{j=1}^{|c|} p_j = 1$. Then, we compute the entropy $H(s_i)$ to identify those regions providing a probability distribution with dominant classes: A region that includes a dominant class will minimize the entropy, while a weakly classified segment with nearly uniform distribution will obtain a high entropy. Our classification consistency measurement, based on the entropy $H(s_i)$ for a region $s_i \in S$, is defined as

$$H(s_i) = -\sum_{j=1}^{|c|} p_j \log p_j, \qquad p_j = \frac{\log P(c=j|s_i)}{\sum_{k=1}^{|c|} \log P(c=k|s_i)}. \tag{5}$$

Considering the set of generated segmentations S and the corresponding consistency measurements $H(S) = \{H(s_1), \ldots H(s_N)\}$, we construct the final image partition as follows: For each pixel q in image I we assign the corresponding segmentation index $i_q^* \in \{1,\ldots,N\}$ by minimizing the obtained entropies over all segments in S that include q:

$$i_q^* = \arg\min_{q \in s_i} H(s_i). \tag{6}$$

These indexes are stored in an image structure and provide the final partition for the CRF stage, that incorporates the contextual knowledge. Figure 2 shows some examples of generated entropy images, where each pixel includes the minimized entropy obtained by the different partitions in S. Assigning the most likely object class $c_q^* = \arg\max_k \log P(c=k|q,s_i)$ provides the temporary classification maps presented in Figure 2.

4 Integrating the Co-occurrence Information

For each test image we generate an individual co-occurrence matrix θ that represents most probable class configurations, *e.g.*, *cars* cannot appear with *water* or *books* do not appear in images covered with *trees*. The semantic context information is obtained by evaluating the leaf nodes, considering the feature representation, at run-time and accumulating the additionally generated co-occurrence relationships for all pixels in an image. Since we accumulate all permutations of possible class appearances during the training, we obtain the final probability matrix by normalizing $\theta(c_i,c_j)$ by row-wise sums [4, 18]. Then, each element in the final co-occurrence matrix θ gives the joint probability that a class c_i appears with the class c_j. In this work, we use the elements of the normalized matrix $\theta(c_i,c_j)$ as pairwise potentials to integrate the semantic context. Moreover, we perform the labeling on a region adjacency

graph, which is faster than optimizing the labels on a full image grid. The energy $E(c)$ in the can now be defined as

$$E(c) = \sum_i D(c_i) + \sum_{i,j} w_{ij} V(c_i, c_j), \qquad w_{ij} = \lambda \frac{2B(s_i, s_j)}{B(s_i) + B(s_j)}, \tag{7}$$

where $D(c_i)$ denotes the data term including the unary potentials according to $D(c_i) = -\log(P(\mathbf{c}|s_i))$ and $P(\mathbf{c}|s_i)$ is the posterior class distribution of a region s_i. To incorporate the region sizes into the minimization process (favoring larger regions), we compute a penalty term w_{ij} according to the normalized amount of common boundary pixels between the regions s_i and s_j. $B(\cdot)$ is a function for computing the number of boundary pixels of a given region. The factor λ controls the influence of the common border and is learned during the training process. The pairwise class potentials $V(c_i, c_j)$ include the contextual knowledge and are computed according to $V(c_i, c_j) = -\log(\theta(c_i, c_j))\delta(c_i \neq c_j)$. In this work we apply the efficient primal-dual strategy of [11] to minimize the energy defined in Equation 7.

5 Experimental Evaluation

In this section we present extensive results of our proposed method. The experiments mainly concentrate on the two versions of the MSRC datasets [21, 22]. However, we also report a classification rate for the challenging VOC2007 [6] image collection. The MSRCv1 includes 240 images with available annotations on the pixel level. For the experiments, we randomly split the dataset into 120 training and 120 test images following the evaluation procedure presented in [20]. The MSRCv2 consists of 532 images, that include 21 different labeled classes. The experiments on this dataset are performed using the suggested training/testing splits in [21, 22] to obtain comparable results. 276 images are used for training and 256 for testing. In all accomplished experiments, we collect the training data considering the available ground truth labeling. Since we aim to train a classifier on the pixel level incorporating a small neighborhood of $n = 21$ pixels, an application of the segmentation methods is not required and, therefore, speeds up and simplifies the data acquisition. The RF is only trained on a subset the local feature patches, which are sparsely collected on a 5×5 image grid. In our experiments we train a forest with $T = 10$ trees and a maximum depth of 15. In each binary splitting node we perform tests to select discriminative features maximizing the information gain with respect to the classes. Each tree is trained on a subset of training data including 50000 feature vectors. Due to unbalanced labeling of the training data, we apply an inverse weighting, similar to [22], taking into account the number of samples for each class to simulate a balanced dataset.. At evaluation time, the local patch descriptor is extracted at each position of the test image and parsed down the tree to obtain the final averaged class distribution.

In a first experiment we evaluate all different stages of our approach on the pixel level and compare the results to state-of-the-art performance. We show the raw results obtained by the RF classifier, the classification rates of grouping the pixels using multiple segmentations, and the classification rates using a CRF formulation for the integration of contextual constraints, thus, allowing to asses the importance of the different steps. Additionally, we report the classification rates for a globally estimated co-occurrence matrix taking into account the pure ground truth data (denoted as CRFg). The obtained results at the different stages are summarized in Table 1. The rates are given in terms of averaged pixel accuracies and mean

	MSRCv2		MSRCv1	
Included stages	pixel-level	class average	pixel-level	class average
RF classifier	55.8 %	42.2 %	76.1 %	72.2 %
RF + MS	69.8 %	57.7 %	83.5 %	80.7 %
RF + MS + CRFg	71.3 %	60.1 %	84.1 %	81.3 %
RF + MS + CRFi	73.7 %	61.8 %	86.8 %	81.8 %

Table 1: Classification accuracies on the MSRCv1 and MSRCv2 dataset. The results are evaluated on pixel level and illustrate the performance of the different stages of our approach: Raw pixel classification integrating the covariance feature into the RF classifier, introducing spatial support using the multiple segmentations (MS), and the post processing step with a CRF stage that includes the contextual information (CRFi denotes the individually generated context information using the structure of the RF, CRFg uses the globally estimated co-occurance matrix).

	MSRCv2		MSRCv1	
Method	pixel-level	class average	pixel-level	class average
Ours	73.7 %	61.8 %	86.8 %	81.8 %
Schroff *et al.* [20]	71.7 %	n.a.	87.2 %	n.a.
Gould *et al.* [9]	76.5 %	n.a.	88.5 %	n.a.
Pantofaru *et al.* [16]	74.3 %	60.3 %	n.a.	n.a.
Lazebnik *et al.* [12]	72.14%	62.8 %	n.a.	n.a.

Table 2: Performance comparison in terms of pixel accuracy and averaged classification of each object class on the MSRCv1 and MSRCv2 dataset, showing the final classification results of state-of-the-art approaches.

class percentages. It can be seen that our feature representation, that directly integrates several cues, results in a reliable initial classification. Moreover, an incorporation of multiple unsupervised segmentations significantly improves the rates. In addition, the integration of individually generated context information gives slightly better rates in terms of classification accuracy than the global estimated co-occurrence matrix.

In [20] rates of 69.7% are reported integrating color, HOGs, and textons, however, using an RF classifier with 30 trees, each with a maximum depth of 20. Recently, Lazebnik *et al.* [12] obtained initial pixel-wise rates of about 53.26% by combining SIFTs, texton filter responses, color and spatial information in a bag-of-word model. The approaches [9, 16] reported slightly better results than ours, however they use a relative location prior or additional spatial information. Table 2 compares the obtained classification to reported rates of selected state-of-the-art approaches. Figure 2 shows some visual results. On the challenging VOC2007 images (422 for training and 210 for testing) our approach obtains an averaged class rate of 29.1%, which is in the range of the 2007 winner approach TKK [5].

Figure 3 shows two representative images and the corresponding, individually at evaluation time generated, co-occurrence information. Considering the face image, it can be seen that the *face* and the *body* class are very likely to occur in this image. In case of the water/bird image, the *water* class is likely to occur with *boat*, *tree*, *sky*, and *bird*. The bright colors denote higher probability for a class co-occurrences. The overall computation of the final classification for an image takes less than 10 seconds on a single-core PC. However, the construction of the multiple image segmentations is the most time consuming part of our approach.

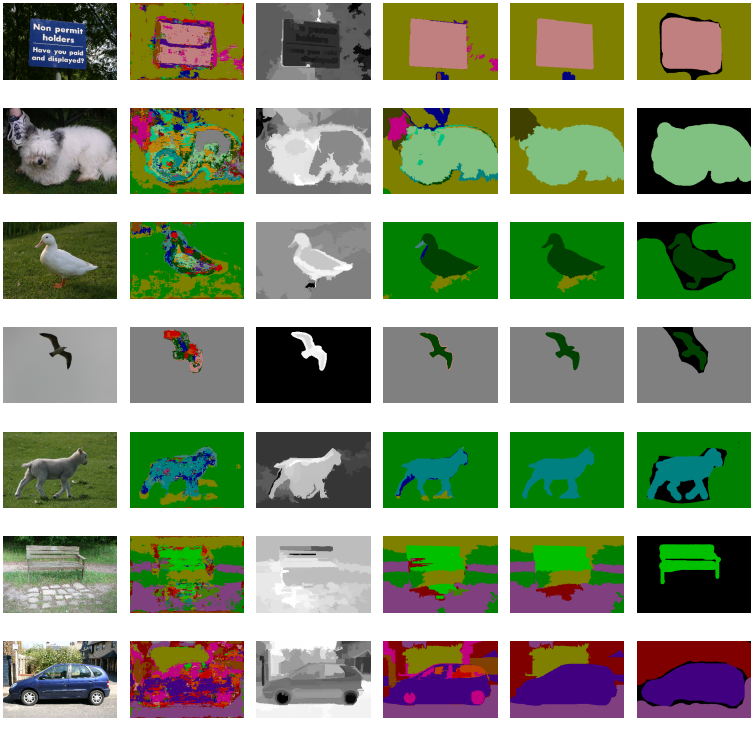

Figure 2: Visual results selected from the classification procedure on the MSRCv2 database including 21 object classes. The first column shows the original color images. The initial results obtained by the RF classification are given in the second column. The third column shows the computed entropies, where a dark color denotes a high classification consistency. Moreover, the semantic labeling results using multiple segmentations, the CRF cleaned final classifications, and the ground truth images are depicted in columns four to six.

6 Conclusion

We have presented an approach for semantic image classification by integrating contextual constraints such as the class co-occurrences into a random forest classification framework. Furthermore, we have illustrated that an initial classification on the pixel level can be used to obtain a reliable region classification. From that, we applied an entropy based measurement to produce a final image partition providing a segmentation with consistently classified regions. A CRF stage then incorporates context knowledge individually generated for each test image. In the experimental section we have demonstrated state-of-the-are performance using our three staged approach. Future work will concentrate on integrating spatial context and on further improving the performance by including the scale of the different object classes into the classification pipeline.

Acknowledgments. This work was supported by the Austrian Science Fund Projects W1209 and P18600 under the doctoral program Confluence of Vision and Graphics, by the FFG projects APAFA (813397) and AUTOVISTA (813395), financed by the Austrian Research

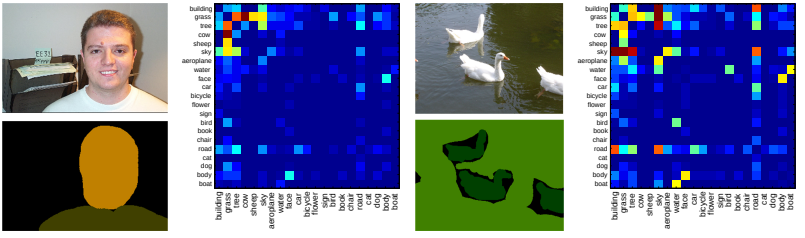

Figure 3: Individually generated co-occurrence matrices that directly correspond to the sample images: Each color-coded entry of the matrices represents the probability of specified class co-occurrences. Bright color denotes a high class probability.

Promotion Agency, and by the Austrian Joint Research Project Cognitive Vision under the projects S9103-N04 and S9104-N04.

References

[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] Leo Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

[3] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[4] H. Jair Escalante, Manuel Montes, and L. Enrique Sucar. Word co-occurrence and markov random fields for improving automatic image annotation. In *Proceedings British Machine Vision Conference*, 2007.

[5] Mark Everingham, L. Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[6] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, September 2004.

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

[8] Carolina Galleguillos, Boris Babenko, Andrew Rabinovich, and Serge Belongie. Weakly supervised object recognition and localization with stable segmentations. In *Proceedings European Conference on Computer Vision*, 2008.

[9] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80 (3):300–316, 2008.

[10] Pushmeet Kohli, Lubor Ladicky, and Philip Torr. Robust higher order potentials for enforcing label consistency. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, 2008.

[11] 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.

[12] Svetlana Lazebnik and Maxim Raginsky. An empirical bayes approach to contextual region classification. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, 2009.

[13] Xi Li, Weiming Hu, Zhongfei Zhang, Xiaoqin Zhang, Mingliang Zhu, and Jian Cheng. Visual tracking via incremental log-euclidean riemannian subspace learning. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, 2008.

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

[15] Frank Moosmann, Eric Nowak, and Frederic Jurie. Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1632–1646, 2008.

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

[17] 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.

[18] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *Proceedings International Conference on Computer Vision*, 2006.

[19] Bryan C. Russell, Alexei A. Efros, Josef Sivic, William T. Freeman, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, 2006.

[20] Florian Schroff, Antonio Criminisi, and Andrew Zisserman. Object class segmentation using random forests. In *Proceedings British Machine Vision Conference*, 2008.

[21] Jamie Shotton, John Winn, Carsten Rother, and Antonio Crimininsi. Textonboost: Joint appearance, shape and conext modeling for muli-class object recognition and segmentation. In *Proceedings European Conference on Computer Vision*, 2006.

[22] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, 2008.

[23] Ranjith Unnikrishnan, Caroline Pantofaru, and Martial Hebert. Toward objective evaluation of image segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):929–944, June 2007.

[24] Jakob Verbeek and Bill Triggs. Scene segmentation with crfs learned from partially labeled images. In *Neural Information Processing Systems*, 2007.

[25] Lin Yang, P. Meer, and D.J. Foran. Multiple class segmentation using a unified framework over mean-shift patches. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, 2007.