Technical Report TR-06-2010
June 2010

Computer Vision and Machine Learning Group
Institute of Numerical Simulation
University of Bonn

Image Segmentation by Discounted Cumulative Ranking on Maximal Cliques

Joao Carreira, Adrian Ion and Cristian Sminchisescu

Computer Vision and Machine Learning Group,
Institute for Numerical Simulation,
Faculty of Mathematics and Natural Sciences, University of Bonn
{carreira, ion, cristian.sminchisescu}@ins.uni-bonn.de

Abstract

We propose a mid-level image segmentation framework that combines multiple figure-ground hypothesis (FG) constrained at different locations and scales, into interpretations that tile the entire image. The problem is cast as optimization over sets of maximal cliques sampled from the graph connecting non-overlapping, putative figure-ground segment hypotheses. Potential functions over cliques combine unary Gestalt-based figure quality scores and pairwise compatibilities among spatially neighboring segments, constrained by T-junctions and the boundary interface statistics resulting from projections of real 3d scenes. Learning the model parameters is formulated as rank optimization, alternating between sampling image tilings and optimizing their potential function parameters. State of the art results are reported on both the Berkeley and the VOC2009 segmentation dataset, where a 28% improvement was achieved.

1 Introduction

Segmenting an image into multiple regions has for long been considered a plausible precursor of many high level visual recognition routines. Indeed, if plausible image regions could be extracted so they would at least partly overlap the projections of visible surfaces in the scene, it would be conceivable that such interpretations can be later lifted to high-level scene percepts by invoking part-based object models and scene consistency rules. This has motivated research into (hierarchical) multipart image segmentations, for which many excellent methods are available [1, 2, 3, 4]. But finding good multipart image segmentations in one step has proven difficult, partly due to the inherently local nature of the grouping process. The competition constraints implicit in various methods make it difficult to integrate scene constraints and mid-level grouping into early computations, and can influence results in ways that do not always correlate with scene properties. Learning segmentation models has also been problematic, partly because of insufficient support for reliable feature extraction and because inference, the inner core of learning, is usually very expensive.

The alternative computational framework we pursue assembles multipart image interpretations by tiling multiple figure-ground image segment hypotheses using mid-level scene constraints. The problem of hypothesis selection and consistent (full) image segmentation is formulated as optimization over sets of maximal cliques, sampled from a graph that connects non-overlapping image segments. By designing and learning clique potentials that encode both intrinsic, unary Gestalt segment properties and pairwise spatial compatibilities that account for plausible configurations of neighboring, spatially non-overlapping segments, we are able to eliminate many implausible image segments and tilings that cannot possibly arise from the projection of surfaces in typical, structured 3d scenes. We show that such a strategy achieves the state of the art in benchmarks like Berkeley and VOC2009.

1.1 Related work

Approaches to image segmentation include normalized cuts [1], mean shift [2] and minimum spanning trees [3]. They are usually computed multiple times, to increase the probability that some of the retrieved segments capture full objects, or their significant parts in images. Another methodology to obtain multiple segmentations is to aggregate in a hierarchy, two well-known examples being multigrid methods [5] and the Ultrametric Contour Maps [4]. The latter achieved state-of-the-art results in a number of challenging segmentation datasets. These algorithms partition the image into a number of regions by using pairwise pixel dependencies. Direct learning is usually targeted at finding the parameters of local affinities [4, 6]. Other techniques work at coarser scales by optimizing over superpixels. This allow features to be computed over a larger spatial support. Ren and Malik [7] learn a classification model to combine superpixels based on their Gestalt properties. Hoiem et al [8] proposed a model that reasons jointly over scene geometry and occlusion boundaries, progressively merging superpixels so as to maximize the likelihood of a qualitative 3d scene interpretation. Instead our goal is complementary: a set of consistent full image segmentation hypotheses, computed based on mid-level Gestalt cues and implicit 3d constraints.

While multi-part image segmentation algorithms are most commonly used, a number of figure-ground methods have been recently pursued. Bagon et al [9] proposed an algorithm that generates figure-ground segmentations by maximizing a self-similarity criterion around a user selected image point. Malisiewicz and Efros [10] showed that good object-level segments could be obtained by merging pairs and triplets of segments from multi-part segmentations, but at the expense of generating also a large quantity of implausible ones. Carreira and Sminchisescu [11] generate a compact set of segments using parametric minimum cuts and learn to score them using region and Gestalt-based features. These algorithms were shown to be quite successful in extracting full object segments, suggesting that a promising research direction is to develop methods that combine multiple figure-ground segmentations (or just segments obtained at multiple scales, potentially from different methods), into plausible full image segmentations. Still missing is a formal multiple hypothesis computational framework for consistent selection (tiling) and learning, which we pursue here. Providing a compact set of multiple hypotheses rather than a single answer is desirable for learning, for high-level, informed processing and for graceful performance degradation.

Organization: In sec. 2 we present our maximal clique formulation framework including both the search procedure and the parameterization of the clique potentials. Sec. 3 describes our ranking-based learning framework that alternates between sampling new tilings (a discrete optimization method) and optimizing the parameters of our clique potentials (a continuous problem) against the test error measure, here the full image segmentation quality. Sec 4 discusses our segment, mid-level unary and pair-wise terms based on Gestalt measures and the statistics of projected boundaries of 3d surfaces, including T-junctions and extremal edges. We show inference and learning statistics as well as experiments on the Berkeley and Pascal VOC 2009 segmentation datasets in sec. 5. We conclude with ideas for future work in sec. 6.

2 Image tiling as sampling maximal cliques

Given a set $\mathcal{H}$ of N segments our aim is to generate several tilings c such that no two segments on c overlap and c has a high score $F_\theta(c)$. Consider for that a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, called the *consistency graph*, where the vertices are the segments in $\mathcal{H}$. Two vertices are connected by an edge if the corresponding segments do not overlap.[1] A *clique* of $\mathcal{G}$, which is a fully connected subgraph of $\mathcal{G}$, corresponds to a set of segments that can form a tiling. A clique is called *maximal*[2] if it is not included into any other clique and hence a larger clique cannot be obtained by adding vertices to it. In our case a maximal clique corresponds to a tiling that cannot be extended using any other segment in $\mathcal{H}$. A *maximum clique* of a graph is a clique with the largest number of vertices. A *maximum weighted clique* is a clique that maximizes the sum of weights associated to its vertices.

[1] While disallowing overlap increases the exposure to imperfect boundary alignments between the available segments, it leads to a dramatic reduction in the solution space and doesn't require additional processing to assign pixels lying on the intersection of overlapping segments.

[2] Also called inclusion maximal clique.

Algorithm 1 *FG-Tiling*$(\mathcal{H}_I, \boldsymbol{\theta}, \boldsymbol{\phi}_u, \boldsymbol{\phi}_p)$ - Discrete optimization for image tilings.

```
Input: Pool of $N$ segments $\mathcal{H}$, weights $\boldsymbol{\theta} = (\boldsymbol{\theta}_u, \boldsymbol{\theta}_p)$, features $\boldsymbol{\phi}_u, \boldsymbol{\phi}_p$.
 1: $\{s_i\}_{i=1...N} \leftarrow$ segments in $\mathcal{H}$ in decreasing order of $\boldsymbol{\theta}_u^T \cdot \boldsymbol{\phi}_u(s_i)$ /* unary terms */
 2: for $i = 1 \ldots N$ do
 3:   $c_i \leftarrow \{s_i\}$ /* initialize clique */
 4:   /* Step 1: sequential greedy heuristic to "build" maximal clique */
 5:   for $j = 1 \ldots n$ do
 6:     if $s_j$ does not overlap any segment in $c_i$ then
 7:       $c_i \leftarrow c_i \cup \{s_j\}$
 8:     end if
 9:   end for
10:   /* Step 2: local search heuristic for solution refinement */
11:   repeat
12:     for all $s' \notin c_i, s'$ not overlapping $s_i$ do
13:       $c' \leftarrow (c_i \setminus \mathcal{O}(s')) \cup \{s'\}$ /* remove segments that overlap $s'$, add $s'$ */
14:       $c' \leftarrow c' \cup \{s_{l1}, s_{l2}, \ldots\}$ /* extend $c'$ to a maximal clique like in lines 5–9 */
15:       if $F_\theta(c') > F_\theta(c_i)$ /* see eq. 1 */ then
16:         $c_i \leftarrow c'$
17:       end if
18:     end for
19:   until convergence
20: end for
Output: Pool of tilings $\mathcal{C} = \cup\{c_i\}$ for the current image ranked in decreasing order of $F_\theta(c_i)$.
```

We formulate the search for tilings as finding *maximal cliques* $c \subseteq \mathcal{G}$ with high potential $F_\theta(c)$

$$F_\theta(c) = \sum_{s_i \in c} \boldsymbol{\theta}_u^T \cdot \boldsymbol{\phi}_u(s_i) + \sum_{s_i \in c, s_j \in c \cap \mathcal{N}(s_i)} \boldsymbol{\theta}_p^T \cdot \boldsymbol{\phi}_p(s_i, s_j) \tag{1}$$

where $\boldsymbol{\phi}_u(s_i)$, $\boldsymbol{\phi}_p(s_i, s_j)$ are feature vectors extracted for, respectively, segment s_i and *image neighbors* s_i, s_j (denoted $s_j \in \mathcal{N}(s_i)$). $\boldsymbol{\theta} = (\boldsymbol{\theta}_u, \boldsymbol{\theta}_p)$ are the corresponding weights learned as mentioned in Section 3.

The problem of finding the *maximum* (weighted) clique of a general graph is known to be both NP-complete and hard to approximate to a given bound [12]. Existing algorithms produce one single solution which equals or approximates the maximum clique. In the weighted case maximization is done only over unary terms associated to vertices. This is different from our case. We desire multiple tilings for each image and the potential of a clique (tiling) depends on both unary and pairwise terms. Enumerating all cliques to find the optimum is not feasible as we deal with many vertices (over 150) and the complexity to enumerate all cliques of size k of a graph with N vertices is $O(N^k k^2)$. Finding a *maximal* clique can be done in linear time in the number of vertices, by starting with one vertex and adding each of the other vertices in some order. But graphs that have a large maximum clique can have maximal cliques of arbitrary small size. To obtain multiple estimates we follow a two step greedy approach: *(*i) starting with each vertex generate a maximal clique; *(*ii) refine each solution using a local search in the space of maximal cliques based on the trained cost function. We generate up to N different tilings, ranked in decreasing order of F_θ. Notice that our approach is based on established strategies to find approximations of the maximum clique (step 1 is known as a *sequential greedy heuristic* and step 2 as a *local search heuristic* [12]). *Algorithm* 1 describes the proposed method.

Complexity: The size of the largest clique that can be formed with a certain vertex is bounded by the degree of this vertex, in our case $d = \deg(s_i)$. If a set $\mathcal{U}$ is kept, containing the segments in $\mathcal{H} \setminus c_i$ which are not overlapping any segment in c_i, the complexity of step 1 is $O(N + d^2)$. Maximum N steps are needed to build $\mathcal{U}$ from the list of sorted segments and d^2 is an upper bound for the loop in step 1 and the verification inside.

Step 2 can be executed in $O(Md(d+N+d^2))$ where M is the maximum number of iterations allowed[3]. The inner loop over all s' is bounded by d as s' must not overlap s_i. Rejecting segments in c' overlapping

[3] In experiments we use $M = 10$.

Algorithm 2 Learning algorithm that estimates parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_u, \boldsymbol{\theta}_p)$.

```
Input: Segments $\mathcal{H}_I$ for the images in the training set $T$, features $\boldsymbol{\phi}_u, \boldsymbol{\phi}_p$, rank $K$.
$\boldsymbol{\theta} \leftarrow (\boldsymbol{\theta}_{u0}, \boldsymbol{\theta}_{p0})$ /* Initialize weights */
for all $I \in T$ do
    $\mathcal{C}_I \leftarrow$ FG-Tiling($\mathcal{H}_I, \boldsymbol{\theta}$) /* extract initial tilings for image $I \in T$ */
end for
repeat
    $\boldsymbol{\theta} \leftarrow \arg\max_\theta \{\sum_{I \in T} S_\theta(\mathcal{C}_I, K)\}$ /* optimization: find $\boldsymbol{\theta}$ that maximizes $S_\theta$ */
    for all $I \in T$ do
        $\mathcal{C}_I \leftarrow$ FG-Tiling($\mathcal{H}_I, \boldsymbol{\theta}, \boldsymbol{\phi}_u, \boldsymbol{\phi}_p$) /* extract new tilings for image $I \in T$ */
    end for
until no improvement
Output: Weights $\boldsymbol{\theta} = (\boldsymbol{\theta}_u, \boldsymbol{\theta}_p)$.
```

with s' is also bounded by d as all segments previously in c' are not overlapping s_i. Finally, extending c' to a maximal clique has the same complexity as step 1, namely $O(N + d^2)$.

Ordering the segments is done only once, thus the complexity for running *FG-Tiling* for all segments $s_i \in \mathcal{H}$ is $O(N \log N + N(N + d^2 + Md(d + N + d^2)))$ where the dominant worst case component is $O(Nd^3)$ if M is fixed. In practice our matlab implementation using $M = 10$ takes on average 20 seconds per image for the BSDS test set.

3 Learning mid-level vision

Assume we are given a set of features $\boldsymbol{\phi}_u(s_i), \boldsymbol{\phi}_p(s_i, s_j)$ computed, respectively, for segments $s_i \in \mathcal{H}$ and pairs of segments $s_i, s_j \in \mathcal{H}$ which are neighbors in the image, i.e. s_i, s_j share a common boundary and do not overlap. We search for the weights $\boldsymbol{\theta} = (\boldsymbol{\theta}_u, \boldsymbol{\theta}_p)$ such that the ranking of tilings induced by F_θ (eq. 1) is as close as possible to the ranking induced by the quality of the tilings with respect to the ground truth.

The learning process alternates between the discrete optimization of tilings, where it runs *FG-Tiling* with the existing parameters $\boldsymbol{\theta}$ to create a new pool of tilings for each of the images in the training set, and a continuous parameter optimization step that finds parameters $\boldsymbol{\theta}$ which maximize an objective function S_θ on the produced tilings, as used for testing: the overlap with ground truth (*Algorithm* 2). Instead of aiming to enforce only the best tiling in the first position, which might be impossible, we design a scoring (with best as special case) that aims at ranking tilings in decreasing order of their quality. For an image I, weights $\boldsymbol{\theta}$, and a pool of tilings $\mathcal{C}_I = \{c_1, c_2, \dots\}$ where c_i is the tiling at rank i when sorting $\mathcal{C}_I$ in decreasing order of the value of $F_\theta(c_i)$, the objective function $S_\theta(\mathcal{C}_I)$ is:

$$S_\theta(\mathcal{C}_I, K) = \sum_{i=1}^{K} w(i) Q(c_i) \tag{2}$$

where $Q(c_i)$ is the quality of c_i measured using the ground truth, $w(i)$ is the weighting of rank i, and K is the rank parameter which determines the constraint we want to enforce (e.g. $K = 1$ for only the best ranked, $K = |\mathcal{C}_I|$ for a full K-ordering). We define $Q(c_i)$ as the average *covering* of c_i with all ground truth segmentations as in [4]. The covering is the sum of overlaps between each individual segment in a ground truth segmentation and the closest segment in a tiling, multiplied by the area of the ground truth segment. $O(s_i, s_g) = |s_i \cap s_g| / |s_i \cup s_g|$ is the standard overlap measure between s_i and s_g [13]. For rank weighting, we use: $w(i) = w(K, i) = 1/[1 + (i-1)/(K-1)]$. This decay is similar to the *Discounted Cumulative Gain* (DCG) [14] that uses a logarithmic reduction factor of the form $1/\log_2(i+1)$. DCG penalizes more aggressively the error in the first ranks. We found this to work slightly less well in our tests.[4]

[4] For an image segment graph $\mathcal{G}$ with N nodes, clique potentials can be used to define a constrained probability distribution over partitions. We can write a Gibbs distribution over cliques as $p_{\boldsymbol{\theta}}(c) = \exp(F_\theta(c)) / \sum_{c \in \mathcal{G}} \exp(F_\theta(c))$, and can learn using ML, with partition functions approximated by summing only over $O(N)$ cliques, computed by *FG-Tiling* (This approach will be presented in an upcoming technical report.). Here we choose a different loss that directly optimizes the

4 Mid-level image descriptors

Our model aims to generate full image tilings that have properties similar to the ones of ground truth segmentations produced by human annotators. We use both unary features inspired by Gestalt properties and pairwise features sensitive to the boundary statistics arising from projections of 3d surfaces, for a total of 46 unary and 22 pairwise features. These features are computed once and do not change during learning and inference. All features are individually normalized to zero mean and standard deviation 1. **Unary Descriptors:** As unary features, we primarily use the ones proposed in [11], that include the amount of **contrast along the boundary** of the segment (8 features), $\boldsymbol{\phi}_u^b(s_i)$, region properties such as position in the image, area and orientation, $\boldsymbol{\phi}_u^r(s_i)$ (18 features), as well as **Gestalt** properties such as convexity and dissimilarity between the segment interior and the rest of the image in terms of intensity and texture, $\boldsymbol{\phi}_u^d(s_i)$ (8 features).

We complemented the unary features in [11] with a novel set of 12 responses quantifying **center-surround dissimilarity**, $\boldsymbol{\phi}_u^l(s_i)$. We define three image strips of width 18, 30 and 42 pixels around each segment. We compute how dissimilar each strip and the segment are according to 4 different local features: hue, rgb, SIFT and textons. For each type of local feature and each strip, dissimilarity is determined as the chi-square distance between the histogram of quantized local features in the strip and in the segment, resulting in the 12 features. The local features are sampled on a regular grid, every 10 pixels. The color histograms use patches 4 and 8 pixels wide, while the SIFT patches are 8 and 18 pixels wide. The textons are the ones used in globalPb [4] quantized into 64 bins. We quantize the other features into 30 bins, with the codebook being obtained *in each image* at test time by k-means.

Pairwise Descriptors: We define a segment neighborhood between pairs of segments sharing a boundary and not overlapping. The occurrence of such pairs is usually non-accidental, particularly in our pool of figure-ground segmentations, because we don't consider the *ground*. Segments that are artifacts of the particular parameter and location constraint that generated them will tend to have few neighbors. Computing this type of neighborhoods can be done robustly by growing all segments by a small amount (4 pixels in our implementation) and then detecting the pairs that overlap. The pairwise features capture the configuration of pairs of segments. We use two sets of pairwise features. The first encodes **pairwise region properties** such as relative area, position and orientation and is simply defined by $\boldsymbol{\phi}_p^r(s_i, s_j) = |\boldsymbol{\phi}_u^r(s_i) - \boldsymbol{\phi}_u^r(s_j)|$ (18 features).

We also employ 4 features which signal occlusion. In ground truth segmentations, neighboring segments often correspond to projections of objects at different depths, which result in distinctive image statistics. These are sufficiently informative even for determining which of the two neighboring regions corresponds to the occluding surface in 3D space, the so called figure-ground assignment problem [15] [16]. The occluding segment usually has a higher convexity coefficient and is often surrounded by the occluded segment. Let $a(s_i)$ be the unary convexity feature in $\boldsymbol{\phi}_u^d(s_i)$. Then the **relative convexity** feature is implemented as $\boldsymbol{\phi}_p^a = |a(s_i) - a(s_j)|$. Let the length of the adjacent boundary between two segments be l_{12}, and the segment perimeters be l_1 and l_2. Then **surroundedness** is defined as $\boldsymbol{\phi}_p^s = |l_1/l_{12} - l_2/l_{12}|$. Another important occlusion features are **t-junctions**, boundary patterns shaped as a T, usually caused by the intersection of the boundaries of two objects in an occlusion relationship. Typically the location of the leg of the T indicates which segment is occluding the other. T-junctions were used in recent approaches to figure-ground assignment, as an energy term for triplets of regions in CRFs [15] [16] [17]. Here we model them directly as a pairwise segment compatibility feature, by measuring the consistency with which the leg of the t-junctions belongs to the same segment, weighted by the quality of the fitting of the junction to a T, as opposed to being Y-like. The feature is defined as $|\boldsymbol{\phi}_p^t = \sum_k [b_i(t_k) - b_j(t_k)]|$, with the sums being over all junctions between the pair of segments. The weighting is $b_i(t) = \exp(-|(\pi/2 - \alpha_t|)$, α_t being the angle formed by the leg of the junction t with the base. When the leg of the junction is on the boundary separating both segments, or the leg is not on the boundary of segment i then $b_i(t)$ is set to 0. Junctions are hard to detect when considering pixel intensities locally, even for humans [18]. But given a pair of neighboring segments this can be done robustly, as illustrated in fig. 4.

The **shading along region borders** was shown to provide information about occlusion in both

overlap measure used during test time. Notice however, our very different use of cliques compared to product expansions in graphical models. Along this path, modeling the nodes as binary variables in a random field would neither produce the semantics we need, nor would necessarily lead to clique consistent inference.

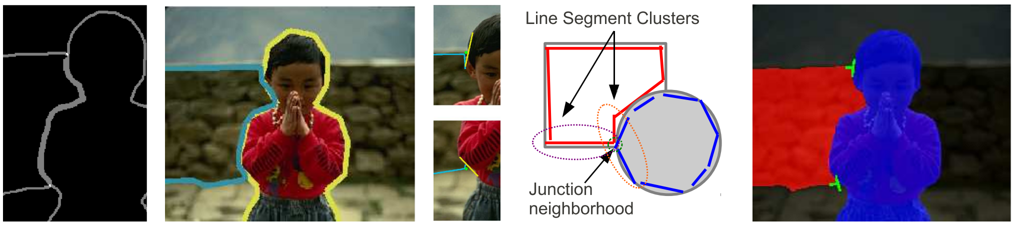

Figure 1: Our T-junction detector works on all pairs of non-overlapping and spatially neighboring segments. In order to detect junctions, we grow the two regions plus their shared background, sum the three binary masks, and find the points in the image where the sum is maximized (first image on the left). These are initial junction points, and are improved by solving a least squares problem minimizing the distance to the closest line segments approximating the boundaries of the two regions (image 2 and 3). To form the base and the leg of the T, these line segments are clustered into two sets based on their orientation, using agglomerative clustering, and a line is fit to each cluster (image 4). The cluster having a line segment endpoint with maximum minimal deviation from the junction along its fitted line is set as the base of the T. The final result is shown on the last image on the right.

computational [19] and psychophysical tests, under the name of *extremal edges* [20]. The phenomenon is explained by the illumination gradient tending to be orthogonal to the boundary, on the occluding side. We implement the gradient orthogonality feature $o(s_i, s_j)$ as in [16] and produce the compatibility feature as $\boldsymbol{\phi}_p^e(s_i, s_j) = |o(s_i, s_j) - o(s_j, s_i)|$. The absolute value is computed because we're not interested here in determining which segment is in front, just in having an occlusion indicator.

5 Experiments

Our inference and learning methods were tested on the Berkeley Dataset (BSDS) [21] and on the Pascal VOC 2009 Segmentation Dataset (VOC2009) [13]. For comparison we show results of the Oriented Watershed Transform Ultrametric Contour Maps using globalPb as contour detector (gPb-owt-ucm) [4].

We generate a pool of segments using the publicly available implementation of Constrained Parametric Min-Cuts (CPMC) [11], which produces nested sets of segments around rectangular seeds on a regular grid with predicted qualities for each segment. Per image an average of 194 segments is generated for the BSDS test set and 156 segments for the VOC2009 validation set. This algorithm was recently shown to produce compact sets of segments that accurately cover ground truth objects.

Fig. 2 shows the evaluation of *FG-Tiling* and two baselines, *Enum-1min* and *Constrained-random*, on the BSDS dataset (see sec. 5). All methods produce maximal cliques i.e. tilings with segments that do not overlap and the cliques cannot be extended using the current pool of segments. For each method the produced tilings are ranked using the scoring function in eq. 1.

Enum-1min is an algorithm that recursively, exhaustively, enumerates maximal cliques until the given time of 1 minute per image is reached and returns the highest scoring N cliques that have been found[5]. Similar to line 1 of *FG-Tiling*, *Enum-1min* first sorts the segments based on $\boldsymbol{\theta}_u^T \cdot \boldsymbol{\phi}_u(s_i)$. During enumeration, it quickly finds one tiling similar to the result of step 1 in *FG-Tiling*. However, within 1 minute, it produces only small variations of the same tiling, as seen also in fig. 2, right. *Constrained-random* is similar to step 1 in *Algorithm* 1 with the difference that in line 1 the order of the segments is randomized. The method gets a few "lucky shots" which explains the quite high values in the plot in fig. 2 left, but overall the average quality of the produced tilings is much lower than the other two methods (23% less than *FG-Tiling* on the test set of BSDS). *FG-Tiling* balances the diversity and quality of the produced tilings to give the best results of all methods.

During learning, for the initial run of *FG-Tiling* we set the weights $\boldsymbol{\theta}_p$ corresponding to the pairwise terms to zero. The weights $\boldsymbol{\theta}_u$ corresponding to the unary terms are set using linear regression s.t. $\boldsymbol{\theta}_u^T \cdot \boldsymbol{\phi}_u(s_i)$ approximates the response $\max_{s_g \in G_I} O(s_i, s_g)$ where G_I is the set of ground truth segments for the image. Parameter optimization is done using a Quasi-Newton method. During this step, the sum

[5] The time of 1 minute given to *Enum-1min* is equal to 3 × the average time of *FG-Tiling* on the BSDS test set. Without the time constraint the algorithm did not finish enumerating cliques after 48 hours on a test image where a pool of $N = 120$ figure-ground segmentations had been used.

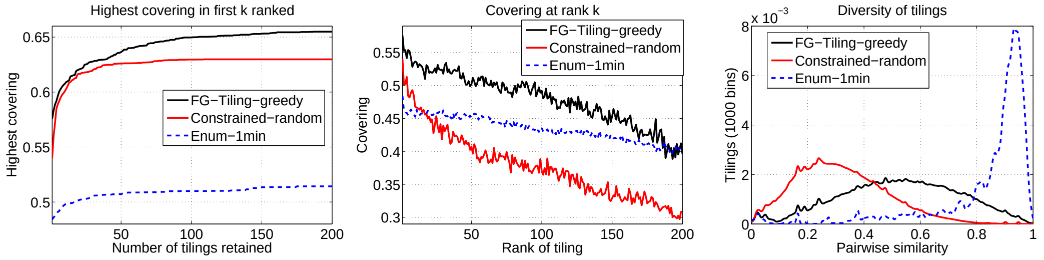

Figure 2: Evaluation of *FG-Tiling*, *Constrained-random*, and *Enum-1min* on the BSDS dataset. *Left*: highest quality $Q(c)$ for number of tilings considered. *Center*: average quality for given rank (if exists). The quality $Q(c)$ was measured with respect to the ground truth using the *covering* measure (see sec. 5 for details). *Right*: histogram of pairwise similarity between produced segmentations.

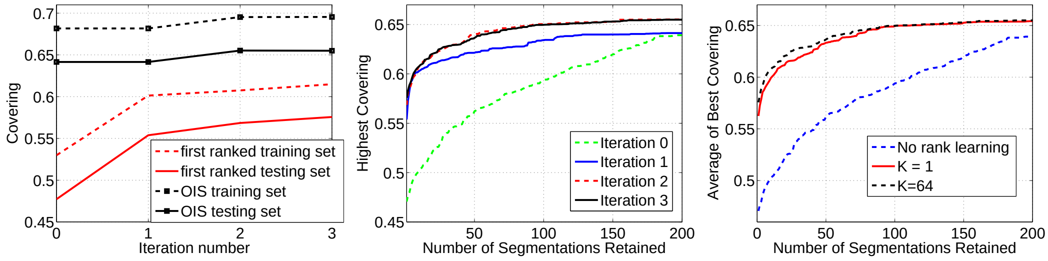

Figure 3: *Left, center*: progress of learning for $K = 64$ rank optimization on the BSDS dataset. *Left*: progress of the first ranked and the highest quality tilings on the training and testing sets. Iteration 0 corresponds to the results with the initial weights $\boldsymbol{\theta}_{u0}, \boldsymbol{\theta}_{p0}$, iteration 1: the same tilings after the first optimization step, iterations 2–3: after new tilings and learned weights. *Center*: highest quality vs. number of segmentations retained on the BSDS test set. *Right*: highest quality vs. number of segmentations retained with no rank learning, learning with rank parameter $K = 1$ and $K = 64$.

of S_θ over all images in the training set and their corresponding pools of tilings is maximized. The first time this step is executed, the initial weight estimates $\boldsymbol{\theta} = (\boldsymbol{\theta}_u, \boldsymbol{\theta}_p)$ required to initialize the search are obtained using linear regression over all tilings produced for the training set. Regression uses targets $Q(c)$ for each tiling c. The inner loop (line 6) needs on average 15 iterations to converge. The outer loop (lines 5–10) saturates after a few iterations (3–4) and both the quality of the first ranked tiling as well as the highest quality over all tilings for each image are maximized. Fig. 3 shows the progress of learning on the Berkeley Segmentation Dataset (BSDS) [21] using $K = 64$ and a comparison of the results: without learning, learning with $K = 1$ and with $K = 64$. We observe that compared to $K = 1$, $K = 64$ produces a slightly better ranking also on the first position, presumably due to the additional constraints from lower ranks.

Table 1 shows results of benchmarks on the test set of BSDS and on the validation set of VOC2009.

BSDS	OIS	First	ODS	BIS
max. possible	0.73	0.73	0.73	1.00
gPb-owt-ucm	0.64	-	0.58	0.74
FG-Tiling	0.64	0.58	-	0.78

VOC2009	OIS	First	ODS	BIS
max. possible	1.00	1.00	1.00	1.00
gPb-owt-ucm	0.58	-	0.45	0.61
FG-Tiling	0.74	0.52	-	0.78

Table 1: Average coverings on the test set of BSDS and on the validation set of VOC2009. The OIS scores for *FG-Tiling* are obtained considering a maximum of 64 respectively 73 tilings per image, which equals the average number of segmentations produced by gPb-owt-ucm [4] on the BSDS test respectively VOC2009 validation sets (notice however that our method uses considerably fewer segments, on average 194 respectively 156 as opposed to 1100 and 1043 in gPb-owt-ucm). If *FG-Tiling* uses the same number of segments but more tilings (on average 176 and 140 in BSDS and VOC2009 respectively), OIS scores of 0.66 respectively 0.76 are obtained.

The values represent average covering scores of ground truth segmentations by the output segmentations. BIS measures the best covering of the ground truth segmentations by individual segments from any segmentation produced by the evaluated method. OIS and ODS have been used in [4] to evaluate the results of gPb-owt-ucm. They have been introduced in the context of hierarchical segmentation, where scale is used to navigate from coarser to finer segmentations. The optimal image scale (OIS) measures for each image the quality of the produced segmentation that best covers the ground truth. The optimal dataset scale (ODS) measures the quality of the segmentations when the same scale is selected for all images. The scale to be evaluated is chosen to maximize the score on the test set. "First" evaluates the results using the predicted best segmentation for each image. "First" is only applicable to our method, since the segmentations from gPb-owt-ucm don't have associated scores to select a single segmentation. ODS is not applicable to our method, as *FG-Tiling* generates independent segmentations. Note that "First" does not use any ground truth information to select the tiling to be evaluated for each image.

The BSDS dataset has multiple ground truth (human) segmentations for each image. To evaluate the quality of a segmentation, the average over all ground truth segmentations for that image is considered. As the provided human segmentations are different, the upper bound for OIS, "First", and ODS on the BSDS test set are 0.73. A score of 1.00 for BIS could be obtained by generating segments that perfectly cover all ground truth segments.

The results obtained by *FG-Tiling* are competitive on BSDS and superior on the VOC2009. Note that the given VOC2009 scores are not using the "segmentation challenge" evaluation which requires recognition, but evaluating the quality of unlabeled segmentations like the method we compare with [4]. The results of gPb-owt-ucm on VOC2009 have been computed by us using the code provided by the authors and are consistent with their published results on VOC2008.

6 Conclusions

We have proposed a mid-level computational learning and inference framework for image segmentation that tiles multiple figure-ground hypotheses into a complete interpretation. The inference problem is formulated as searching for high-scoring maximal cliques in a graph connecting non-overlapping putative figure/ground hypotheses. Clique potentials are based on both intrinsic Gestalt segment quality and compatibilities among neighboring image segments, as derived from statistics of 3d scene boundaries. Learning is formulated as optimizing the ranking of the best-K hypotheses, directly on the testing error, measuring the overlap between image tilings and the ground truth human annotations. We have empirically analyzed the performance of our learning and inference components and have shown that these achieve state of the art results in the Berkeley and the VOC2009 segmentation benchmarks. In the latter the proposed method improves on the state-of-the-art by 28% when considering the full set of generated tilings, and by 16% for the predicted best tiling. In future work we plan to combine segmentation and partial recognition in order to be able to interpret images that contain both familiar and unknown objects.

References

[1] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

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

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

[4] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[5] E. Sharon, M. Galun, D. Sharon, R. Basri, and A. Brandt. Hierarchy and adaptivity in segmenting visual scenes. *Nature*, 442(7104):719–846, 2006.

[6] T. Cour, N. Gogin, and J. Shi. Learning spectral graph segmentation. In *IEEE International Conference on Artificial Intelligence and Statistics*, 2005.

[7] X. Ren and J. Malik. Learning a classification model for segmentation. *IEEE International Conference on Computer Vision*, 2003.

[8] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007.

[9] S. Bagon, O. Boiman, and M. Irani. What is a good image segment? a unified approach to segment extraction. In *European Conference on Computer Vision*, 2008.

[10] T. Malisiewicz and A. Efros. Improving spatial support for objects via multiple segmentations. In *British Machine Vision Conference*, 2007.

[11] J. Carreira and C. Sminchisescu. Constrained Parametric Min-Cuts for Automatic Object Segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[12] I. Bomze, M. Budinich, P. Pardalos, and M. Pelillo. *Handbook of Combinatorial Optimization*, chapter The Maximum Clique Problem, pages 1–74. Kluwer Academic Publishers, 1999.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results. http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

[14] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[15] X. Ren, C. Fowlkes, and J. Malik. Figure/ground assignment in natural images. In *European Conference on Computer Vision*, 2006.

[16] I. Leichter and M. Lindenbaum. Boundary ownership by lifting to 2.1d. In *IEEE International Conference on Computer Vision*, 2009.

[17] D. Hoiem, A. Stein, A. A. Efros, and M. Hebert. Recovering occlusion boundaries from a single image. In *IEEE International Conference on Computer Vision*, 2007.

[18] Josh McDermott. Psychophysics with junctions in real images. *Journal of Vision*, 2(7):131–131, November 2002.

[19] P. Huggins, H. Chen, P. Belhumeur, and S. Zucker. Finding folds: On the appearance and identification of occlusion. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2001.

[20] T. Ghose and S. Palmer. Surface convexity and extremal edges in depth and figure-ground perception. *Journal of Vision*, 5(8):970–970, September 2005.

[21] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE International Conference on Computer Vision*, 2001.

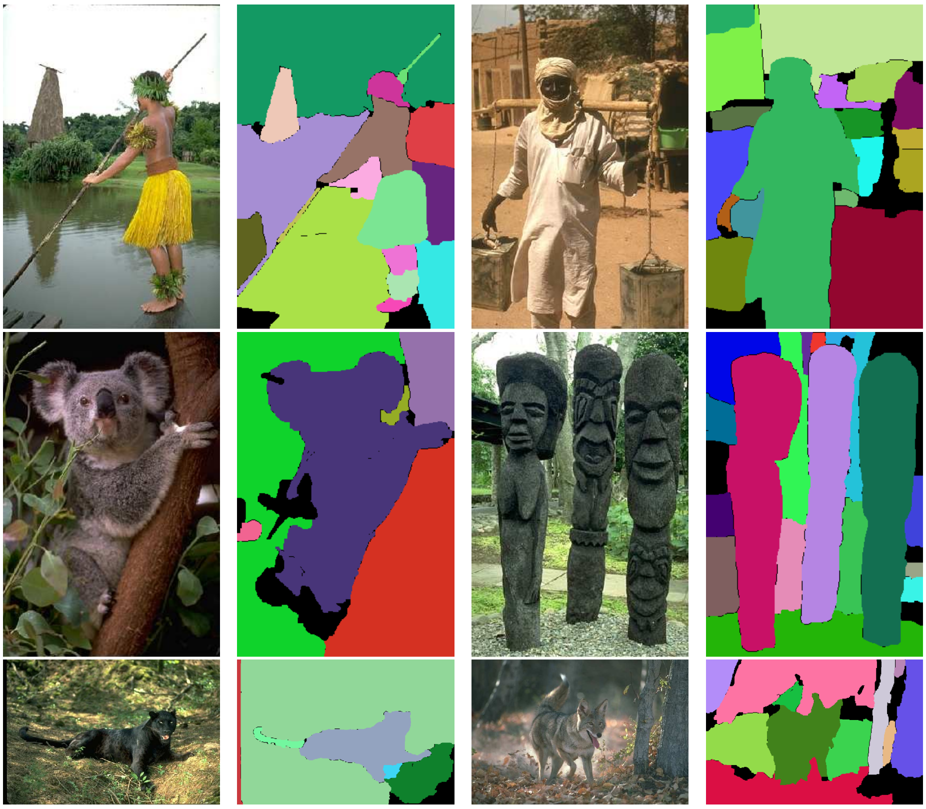

Figure 4: *(Best viewed in color)* Images from the test set of BSDS, accompanied by our highest quality tiling. Areas of the image not covered are colored in black.

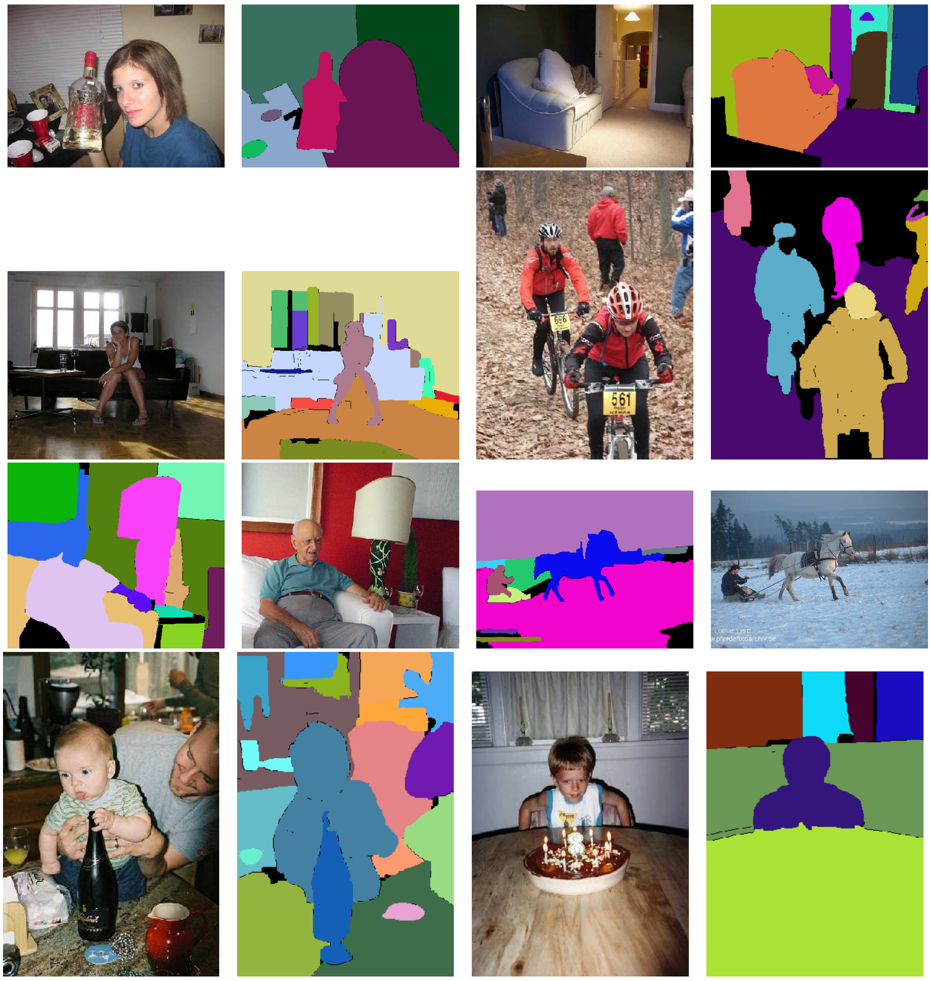

Figure 5: *(Best viewed in color)* Images from the validation set of VOC2009, accompanied by our highest quality segmentation. Non covered areas of the image are colored in black. We colored disconnected segments with the same color, when they have strong occlusion cues with a same third segment, and well as similar color and texture.