To appear in 13th International Conference on Computer Vision, 6-13 November 2011, Barcelona, Spain.

# Image Segmentation by Figure-Ground Composition into Maximal Cliques

**Adrian Ion**[*], **Joao Carreira**, **Cristian Sminchisescu**
Computer Vision and Machine Learning Group, Institute for Numerical Simulation,
Faculty of Mathematics and Natural Sciences, University of Bonn
{ion, carreira, cristian.sminchisescu}@ins.uni-bonn.de

## Abstract

*We propose a mid-level statistical model for image segmentation that composes multiple figure-ground hypotheses (FG) obtained by applying constraints at different locations and scales, into larger interpretations (tilings) of the entire image. Inference is cast as optimization over sets of maximal cliques sampled from a graph connecting all non-overlapping figure-ground segment hypotheses. Potential functions over cliques combine unary, Gestalt-based figure qualities, and pairwise compatibilities among spatially neighboring segments, constrained by T-junctions and the boundary interface statistics of real scenes. Learning the model parameters is based on maximum likelihood, alternating between sampling image tilings and optimizing their potential function parameters. State of the art results are reported on the Berkeley and Stanford segmentation datasets, as well as VOC2009, 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 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 representations 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) multiregion image segmentation (as opposed to binary or figure-ground image segmentation), for which many methods are available [25, 8, 12, 1]. But finding good multiregion 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 produce results (segments) that do not always correlate with the image projection of the objects in the scene. Learning segmentation models has also been problematic not only because of insufficient support for reliable feature extraction, but also because inference, the inner core of learning, is usually very expensive. In this paper, we propose a statistical model that assembles larger scope image interpretations by selecting subsets of hypotheses from a bag of multiple figure-ground segments, based on mid-level scene constraints. The problem of image segmentation is formulated as optimization over sets of max-

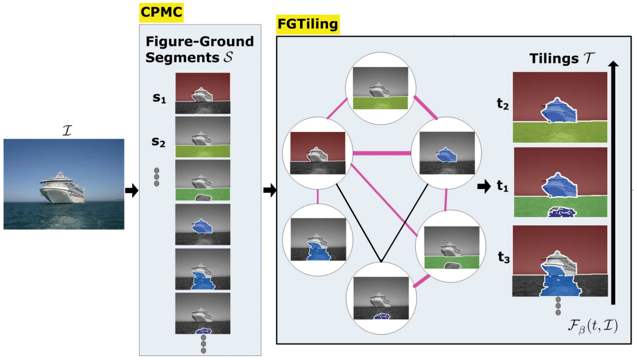
CPMC
Figure-Ground Segments $\mathcal{S}$
$s_1$
$s_2$
$\mathcal{I}$
FGTiling
Tilings $\mathcal{T}$
$t_2$
$t_1$
$t_3$
$\mathcal{F}_\beta(t, \mathcal{I})$

Figure 1. (Best viewed in color) Overview of our segmentation method. Given a bag $\mathcal{S}$ of segments $s_i$ for image $\mathcal{I}$ (figure-ground segmentations obtained using CPMC [7], with *figure* segments selected), we generate different compositions (tilings) of the image from those segments. The problem is formulated as search for maximum weighted cliques $t_i$ in a *consistency graph* that has segments $s_i$ in the nodes and edges between any two segments that do not spatially overlap. The quality (weight) of a clique is given by $\mathcal{F}_\beta(t, \mathcal{I})$ and combines both the intrinsic mid-level quality of segments (object-like regularity like convexity, smoothness of boundary, symmetry) and their mutual compatibility (junction structure, extremal edges) computed over the segment *dependency graph*—a subgraph of the *consistency graph* (edges shown in pink, strength by compatibility) connecting only spatially neighboring segments that share a boundary. Notice that any tiling is made of segment subsets in $\mathcal{S}$, and may induce *residual regions*—image pixels not assigned to any of the segments in $\mathcal{S}$ included in the composition of that tiling. Image segmentation=tiling+residual regions.

[*] Current affiliation: PRIP, Vienna University of Technology & Institute of Science and Technology Austria.

imal cliques, sampled from a graph that connects all non-overlapping figure-ground image segments. Each clique is a possible image segmentation composed of subsets of the figure-ground segments in the bag. 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 real scenes. We show that such a strategy achieves the state of the art in benchmarks like Berkeley, Stanford, and VOC2009. The framework we pursue is depicted in fig.1.

### 1.1. Related work

Approaches to image segmentation include normalized cuts [25], mean shift [8] and minimum spanning trees [12]. 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 a hierarchy, two well-known examples being multigrid methods [24] and the Ultrametric Contour Maps [1]. The latter achieved very good results in a number of difficult datasets. These algorithms partition the image into a number of regions by using pairwise pixel dependencies, with learning focused on local affinities [1, 9]. Other techniques work at coarser scales by optimizing over superpixel combinations. This allows features to be computed over a larger spatial support. Ren and Malik [23] learn a model that combines superpixels based on their Gestalt properties. Hoiem *et al*. [15] proposed a model over scene geometry and occlusion boundaries, progressively merging superpixels so as to maximize the likelihood of a qualitative 3d scene interpretation.

While multiregion image segmentation algorithms are most commonly used, a number of figure-ground methods have been recently pursued. Bagon *et al*. [2] proposed an algorithm that generates figure-ground segmentations by maximizing a self-similarity criterion around a user selected image point. Malisiewicz and Efros [19] showed that segments with good object overlap could be obtained by merging pairs and triplets of segments from multiregion segmentations, but at the expense of generating also a large quantity of implausible ones. Carreira and Sminchisescu [7] generate a compact set of segments using parametric minimum cuts and learn to score them using region and Gestalt-based features (similar ideas have also been recently pursued by [10]). 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 interpretations.[1] Still missing is a formal multiple hypothesis probabilistic computational framework for consistent segment composition (tiling) and learning, which we pursue here. Providing a compact set of multiple hypotheses rather than a single answer is desirable for learning and for graceful performance degradation.

## 2. Segmentation by Composition

In most segmentation methods, partitions of the image into multiple regions are obtained, with segments defined as sets of pixels grouped together by the algorithm. While in such cases, a segmentation begets a set of segments, a segment can also be defined independently of a segmentation. In this work we take this latter approach and decouple the process of computing the segments from the one of obtaining a multiregion image segmentation.

Consider an image $\mathcal{I}$ and a set $\mathcal{S} = \{s_1, \ldots, s_N\}$ of $N$ figure-ground segments obtained by a segmenter that computes independent solutions by applying constraints at different locations and scales in the image. In our case, the segments are obtained using the publicly available Constrained Parametric Min Cuts method (CPMC), but the composition model we present applies just as well to unorganized collections of segments bagged from multiple layers of a single image segmentation method, or to collections obtained by different algorithms. In all cases, some of the segments in $\mathcal{S}$ may spatially overlap.

A *tiling t*, for image $\mathcal{I}$, is a set $t \subseteq \mathcal{S}$ such that: *(i)* no two segments in $t$ overlap [2], and *(ii)* $t$ cannot be extended using any segments in $\mathcal{S}$ while preserving property *(i)*. Property *(i)* says that a tiling needs to be consistent, whereas property *(ii)* ensures it is locally maximal, hence it cannot be extend using another segment in $\mathcal{S}$ while preserving consistency.

The set of all tilings $\mathcal{T}$ of image $\mathcal{I}$, called the *tiling set*, is a subset of the power set of $\mathcal{S}$ and represents all valid (non-overlapping) compositions constructed using segments in $\mathcal{S}$. A full image segmentation is made of a tiling (subset of segments in $\mathcal{S}$), as well as *residual regions* consisting of those image pixels not assigned to any segment included in the composition of that particular tiling. *Residual regions* can either be created as new segments outside

[1] In recent work overlapping with the one presented here (registered by our earlier TR[6]), Brendel and Todorovic[5] pursue an independent set approach to combining segments. This is equivalent to the maximum clique formulation we propose, although we provide a strictly more general inference method that accommodates weights on both vertices and edges, and additionally, a statistical segmentation model and a learning scheme.

[2] The segments are defined by the contained pixels and have fixed positions in the image – they cannot be moved like puzzle pieces. Moreover, while disallowing overlap increases the exposure to imperfect boundary alignments between segments selected in any single tiling, it leads to a dramatic reduction in the solution space and does not raise additional issues with assigning pixels lying on segment intersections.

$\mathcal{S}$, or arbitrarily assigned to segments in the particular tiling.

**Probability model over tilings:** Among the set of tilings for an image, some consist of segments with better intrinsic quality (e.g. object like regularity) as well as mutual compatibility (junctions, boundary structure, etc), than others. It is therefore natural to define a probability distribution over tilings:

$$p_\beta(t; \mathcal{I}) = \frac{1}{Z_\beta(\mathcal{I})} \exp\left(\mathcal{F}_\beta(t, \mathcal{I})\right) \tag{1}$$

with $Z_\beta(\mathcal{I}) = \sum_{t' \in \mathcal{T}} \exp(\mathcal{F}_\beta(t', \mathcal{I}))$ the normalizer or partition function. $\mathcal{F}_\beta(t, \mathcal{I})$ can be defined as interaction over unary and pairwise terms:

$$\mathcal{F}_\beta(t, \mathcal{I}) = \sum_{s_i \in t} \Phi^t(s_i, \beta_u) + \sum_{s_i \in t} \sum_{s_j \in t \cap \mathcal{N}^t_{s_i}} \Psi^t(s_i, s_j, \beta_p) \tag{2}$$

with $\Phi^t$ and $\Psi^t$ unary and pairwise potential functions, and $\mathcal{N}^t_{s_i}$ the image neighborhood i.e. $\mathcal{N}^t_{s_i} = \{s_j \in \mathcal{S} \mid s_i, s_j$ share a boundary and do not overlap$\}$. The parameters of the model are specified by the vector $\beta = [\beta_u^\top \beta_p^\top]^\top$. The unary and pairwise terms are linear in the parameters, e.g. $\Phi^t(s_i, \beta_u) = \beta_u^\top \Phi^t(s_i)$.

**MAP Integer Programming Formulation:** Given the segments $\mathcal{S}$, computing a tiling that maximizes $\mathcal{F}_\beta(t, \mathcal{I})$ under consistency properties *(i)* and *(ii)*, is given by the following integer program:

$$\operatorname*{argmax}_{\mathbf{y}} \sum_{s_i \in \mathcal{S}} \mathbf{y}_i \Phi^t(s_i, \beta_u) + \sum_{s_i \in \mathcal{S}} \sum_{s_j \in \mathcal{N}^t_{s_i}} \mathbf{y}_i \mathbf{y}_j \Psi^t(s_i, s_j, \beta_p) \tag{3}$$

$$\begin{aligned} \text{s.t. } & \mathbf{y}_i \in \{0, 1\}, \forall i \in \{1, \dots, |\mathbf{y}|\} \\ & \sum_i \sum_j \mathbf{y}_i \mathbf{y}_j |s_i \cap s_j| = 0 \\ & \mathbf{y}_i + \sum_j (1 - \mathbf{y}_i) \mathbf{y}_j |s_i \cap s_j| > 0, \forall i \in \{1, \dots, |\mathbf{y}|\} \end{aligned}$$

where $\mathbf{y}_i$ are binary variables indicating the presence of a segment in the tiling, and $|s_i \cap s_j|$ is the number of pixels in the intersection between segments $s_i, s_j$. The first constraint ensures that the variables $\mathbf{y}$ are binary (a segment $s_i$ included in the tiling is encoded as $\mathbf{y}_i = 1$). The second constraint enforces that no pair of selected segments $s_i, s_j$ overlap. The final constraint ensures that an extension of the tiling by adding one more segment in the bag is not possible.

Without the third constraint, or in the case where all potentials $\Phi^t, \Psi^t$ are strictly positive, the problem (3) is equivalent to the maximum weighted clique problem (MWC) in a special structured graph. Consequently, we first consider methods designed for the MWC. We then ensure the third constraint is satisfied, through post-processing.

### 2.1. Tilings as Weighted Maximal Cliques

In this section we present approximations to (2) and (3), based on searching for high scoring maximal cliques in a special constructed graph.

A *clique* of a graph is a subset of its vertices and edges such that any two vertices are connected by an edge. The *weight of a clique* is the sum of the weights of the contained vertices and edges. A clique is called *maximal* if it is not included into any other clique, hence a larger clique cannot be obtained by adding vertices to it. A *maximum weighted clique* (MWC) is a clique that maximizes its weight. A *maximum clique* of a graph is a clique with the largest number of vertices, a special case of the MWC where the vertices have unit weight and edges have zero weight.

Consider a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, called the *consistency graph*, where the vertices are the segments $\mathcal{V} = \mathcal{S}$, weighted by $\Phi^t(s_i, \beta_u)$. Two vertices are connected by an edge if the corresponding segments do not spatially overlap. Edges are weighted by $\Psi^t(s_i, s_j, \beta_p)$ if $s_j \in \mathcal{N}^t_{s_i}$, and 0 otherwise. Notice that $\mathcal{G}$ extends the spatial neighborhood graph induced by $\mathcal{N}^t_{s_i}$ (the *dependency graph*) by adding edges between all non-overlapping pairs of segments to obtain the *consistency graph*.

Finding the maximum (weighted) clique is NP-complete and hard to approximate to a given bound [3]. Different greedy approximations—either continuous or discrete—have been proposed to obtain local solutions. Ultimately, the best performing strategy may depend on the application. In our case, we deal with the most general case of graphs with weights both on vertices and on edges. Moreover, to estimate the partition function for maximum likelihood learning, we need to be able to sample a number of plausible solutions for the current set of parameters, not just obtain one local MAP configuration. Enumerating all cliques grows exponentially in the number of graph vertices. We use in the order of $10^2$ segments (nodes) per image, hence exhaustive search would not be feasible. In the next section we explore (continuous) relaxations and develop novel discrete optimization methods to compute weighted maximal cliques in graphs. Sec. 5 provides comparisons of these methods in complex image segmentation datasets like Berkeley, Stanford and VOC2009.

**Relaxations, vertex-weighted case.** Several relaxations have been proposed to solve the MWC problem for graphs with either their vertices or their edges weighted, but not both. In this section we discuss the more studied vertex-weighted case ($\Psi^t = 0$). The straight-forward relaxation of the MWC is a *linear program*, where the constraints $\mathbf{y}_i \in \{0, 1\}$ are replaced with $\mathbf{y}_i \in [0, 1]$. The program can be solved exactly, but in many cases the solution to the original problem cannot be reconstructed from the continuous one. Indeed we observe a similar behavior for our problem, with most continuous estimates of $\mathbf{y}_i$ stuck at $0.5$.

In Sec. 5 we show additional results based on quadratic relaxations that use the Comtet class of matrices [4]. The advantage of this formulation is that every local optimum of the quadratic program corresponds to a maximal clique of the original graph, which can always be reconstructed from the continuous solution. We study two optimization methods: replicator dynamics [4] and a standard quadratic programming solver with linear constraints. This relaxation requires strictly positive weights, which we obtain by setting all non strictly positive weights to a small $\epsilon > 0$.

**Discrete Search (FGTiling), general case.** Finding a *maximal* clique can be done in linear time in the number of vertices, by starting with one vertex and attempting to add each of the other vertices in some order. But graphs that have a large maximum clique can have maximal cliques of arbitrarily small size. To reduce the chance of identifying valid but low quality cliques, we generate multiple maximal cliques, one expanded from each vertex $s_i$ in the graph.

Existing discrete approximation algorithms produce a single solution which 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 need multiple tilings for each image and the potential of a clique (tiling) depends on both unary and pairwise terms.

To obtain multiple estimates we propose a *two step approach* which we call *FG-Tiling* (as we use figure-ground segments for tiling, but our algorithm can use any bag of segments): *Step 1:* generate a maximal clique starting at each vertex; *Step 2:* refine each solution using local search in the space of maximal cliques based on the cost function $\mathcal{F}_\beta$. We generate up to $|\mathcal{S}|$ different tilings (repeated tilings are removed), ranked in decreasing order of $\mathcal{F}_\beta$. Our joint approach for the general case (arbitrary sign weights on both vertices and edges) exploits ideas shown to be useful in the (single) node/vertex weighted case (*Step 1* is known as a *sequential greedy heuristic* and *Step 2* as a *local search heuristic*). *Algorithm* 1 describes the proposed method in detail. Notice that lines 5–7, 12 ensure that the third constraint in problem (3) is satisfied.

**Complexity of *FG-Tiling*.** 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 $d = \deg(s_i) < N, N = |\mathcal{S}|$. If a set $\mathcal{U}$ is kept containing the segments in $\mathcal{S} \setminus t_i$, which do not overlap any segment in $t_i$, 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. The inner loop over all $s_k$ is bounded by $d$ as $s_k$ must not overlap $s_i$. Rejecting segments in $t'$ overlapping with $s_k$ is also bounded by $d$ as all segments previously in $t'$ are not overlapping $s_i$. Finally, extending $t'$ to a maximal clique has the same complexity as step 1, namely $O(N+d^2)$.

**Algorithm 1** *FG-Tiling*$(\mathcal{S}, \beta)$ - Discrete optimization for tilings (weighted maximal cliques) of image $\mathcal{I}$ under (2).

**Input**: Pool of segments $\mathcal{S}$, weights $\beta = [\beta_u^\top \beta_p^\top]^\top$.

```
1:  ${s_i}_{i=1...|\mathcal{S}|}$ ← segments in $\mathcal{S}$ in decreasing order of
    $\Phi^t(s_i, \beta_u)$ /* order based on unary potentials */
2:  for $i = 1 \ldots |\mathcal{S}|$ do
3:    $t_i \leftarrow \{s_i\}$ /* initialize clique */
4:    /* Step 1: sequential greedy heuristic */
5:    for $j = 1 \ldots |\mathcal{S}|$ do
6:      if $s_j$ does not overlap any segment in $t_i$ then
7:        $t_i \leftarrow t_i \cup \{s_j\}$
8:    /* Step 2: local search heuristic */
9:    repeat
10:     for all $s_k \in \mathcal{S} \setminus t_i, |s_k \cap s_i| = 0$ do
11:       $t' \leftarrow (t_i \setminus OV(s_k)) \cup \{s_k\}$ /* remove segments
          in $\mathcal{S}$ that overlap $s_k$ (set $OV(s_k)$), add $s_k$ */
12:       $t' \leftarrow t' \cup \{s_{l1}, s_{l2}, \ldots\}$ /* extend $t'$ to a maximal
          clique like in lines 5–7 */
13:       if $\mathcal{F}_\beta(t', \mathcal{I}) > \mathcal{F}_\beta(t_i, \mathcal{I})$ /* see eq. 2 */ then
14:         $t_i \leftarrow t'$
15:   until convergence
```

**Output**: Pool of tilings: $\cup\{t_i\}$ for image $\mathcal{I}$, ranked in decreasing order of $\mathcal{F}_\beta(t_i, \mathcal{I})$.

Ordering the segments is necessary only once, and takes $O(N \log N)$. The complexity of *FG-Tiling* given segments $s_i \in \mathcal{S}$ 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 with $M = 10$ takes on average 8s per image for the BSDS test set.

## 3. Learning Mid-level Vision

An important problem for segmentation is learning the model parameters: the plausibility of segments and their mutual compatibility. We learn the parameters $\beta$ (1) using Maximum Likelihood, to maximize $\log p_\beta(t^g; \mathcal{I})$, where $t^g$ is the ground truth segmentation for image $\mathcal{I}$. Computing the partition function $Z_\beta$ is intractable because the number of possible tilings $t'$ is bounded above by the power set of the set (bag) of segments $\mathcal{S}$. Here we use an estimate of $Z_\beta$ obtained by summing only over tilings sampled using *FG-Tiling* (sec 2.1).

At iteration $i$, learning alternates between discrete optimization for tilings, where it runs *FG-Tiling* with the existing parameters $\beta$ to obtain new tilings $\mathcal{T}^i$, and a continuous optimization step that estimates parameters $\beta$ which maximize the probability of the ground truth segmentations, based on summing the tilings in $\mathcal{T}^i$ to estimate $Z_\beta(\mathcal{I})$. For initialization we set the pairwise parameters $\beta_p$ to zero. The weights $\beta_u$ corresponding to the unary terms are initialized using linear regression s.t. $\Phi^t(s_i, \beta_u)$ approximates the response $\max_{g \in G_\mathcal{I}} O(s_i, g)$ where $G_\mathcal{I}$ is the set of ground truth segments for image $\mathcal{I}$ and $O(s_i, g) = |s_i \cap g|/|s_i \cup g|$ is the standard overlap measure between two regions $s_i$ and $g$ [11]. The continuous optimization for $\beta$ is performed using a quasi-Newton method. Experimental details on learning and its effectiveness appear in sec. 5.

## 4. Mid-level image descriptors

We use both unary features inspired by Gestalt properties and pairwise features sensitive to the boundary statistics arising from projections of real scene surfaces: 46 unary and 22 pairwise features. These features are computed once for individual segments and spatially neighboring segment pairs (that share boundaries and do not overlap), and do not change during learning and inference. Features are individually normalized to 0 mean and standard deviation 1.

**Unary Descriptors $\Phi^t(s_i, \beta_u)$:** We primarily use features proposed in [7], that include the amount of *contrast along the boundary* of the segment (8 features), region properties such as position in the image, area and orientation (18 features), as well as *Gestalt* properties such as convexity and intensity and texture dissimilarity between the segment interior and the rest of the image (8 features).

We complemented the unary features in [7] with a novel set of 12 responses quantifying the *center-surround dissimilarity*. We define three image strips of width 18, 30 and 42 pixels around each segment. We compute how dissimilar each strip and the segment interior are according to 4 different local features: hue, rgb, SIFT and textons. The idea is to capture how far the segment boundary is from some image discontinuity. This allows the model to prefer, among imperfect segments, those closer to discontinuities. 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 inside 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[1] quantized into 64 bins. We quantize the other features into 30 bins, with codebook obtained *in each image* at test time by k-means.

**Pairwise Descriptors $\Psi^t(s_i, s_j, \beta_p)$:** For each segment, its dependency neighborhood $\mathcal{N}^t_{s_i}$ (over which pairwise features are defined) consists of all 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 for each, we discard the *ground*, and only keep the *figure*. Computing this type of neighborhoods can be done robustly by growing all segments by a small amount (4 pixels in our implementation), then detecting pairs that start overlapping. We use two sets of pairwise features to capture the configuration of pairs of segments. The first encodes *joint region properties* such as relative area, position and orientation and is simply defined by $|\Phi^t_r(s_i) - \Phi^t_r(s_j)|$ (18 features).

We also employ 4 features to signal occlusion boundaries, which often correspond to desired segment boundaries. Objects at different depths result in distinctive image statistics, that 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 [22, 18]. The occluding segment usually has a higher convexity coefficient and is often surrounded by the occluded segment. Let $a(s_i)$ be our unary convexity feature. The *relative convexity* feature is then implemented as $|a(s_i) - a(s_j)|$. 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 $|l_1/l_{12} - l_2/l_{12}|$. Another important occlusion features are *T-junctions*, structures shaped as a T, usually caused by the boundary intersection 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 [22, 18, 16]. 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 junction fit to a T, as opposed to a Y shape. The feature is defined as $|\sum_k [b_i(t_k) - b_j(t_k)]|$, with sums over all junctions among the pair of segments. The weighting is $b_i(t) = \exp(-|(\pi/2 - \alpha_t|)$, $\alpha_t$ being the angle formed by the leg of the junction 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 difficult to detect when considering pixel intensities locally, even by humans [21]. But given a pair of neighboring segments this can be done robustly. Our procedure is illustrated in fig. 2.

The *shading along region borders* (*extremal edges* [13]) was shown to provide information about occlusion in both computational [17] and psychophysical tests. 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 [18] and produce the compatibility feature as $|o(s_i, s_j) - o(s_j, s_i)|$. The absolute value is computed because we are only interested in an occlusion indicator, not in exact depth ordering.

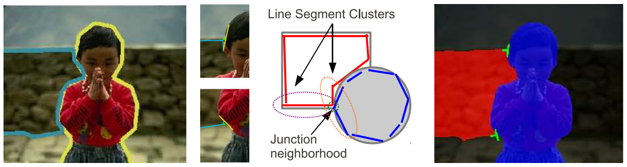

Figure 2. (Best viewed in color). Our T-junction detector works on all pairs of non-overlapping and spatially neighboring segments. In order to detect junctions, we grow the two regions plus their shared background (image complement), add their masks and find the pixels where the sum is maximized. These 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 1 and 2). To form the base and the leg of the T, these line segments are clustered into two sets based on their orientation, and a line is fit to each cluster (image 3). The cluster having a line segment endpoint closest to the junction is set as the leg of the T. The final result is shown on the last image on the right.

## 5. Experiments

Our inference and learning methods were tested on the Berkeley Segmentation Dataset (BSDS) [20], on the Stanford Background Dataset (Stanford) [14] and on the Pascal VOC 2009 Segmentation Dataset (VOC2009) [11]. The BSDS consists of 200 training and 100 test images, each having several ground truth segmentations at the level of objects, object parts and background regions, annotated by multiple subjects. Stanford has 715 outdoor images, and offers, besides semantic and geometric annotations, the annotation called *Regions*, consisting of a single ground truth segmentation into individual object and background regions (5-fold cross-validation is used to report results on Stanford). The VOC2009 has only annotation of individual objects from 20 predefined categories and contains 749 and 750 images in the training and validation sets, respectively. Example ground truth segmentations from these challenging datasets can be found in fig. 5.

To evaluate the similarity between a ground truth segmentation $G$ and an automatically produced segmentation $S$, we use the *covering* measure $C(G,S) = \frac{1}{|\mathcal{I}|}\sum_{g\in G}|g|\max_{s_i\in S} O(s_i,g)$, where $|\mathcal{I}|$ denotes the number of pixels in the image, and $|g|$ the number of pixels in segment $g$. $O(s_i,g) = |s_i \cap g|/|s_i \cup g|$ is the overlap measure between $s_i$ and $g$ [11]. As baseline we use the Oriented Watershed Transform Ultrametric Contour Maps[1] with globalPb as contour detector (gPb-owt-ucm). gPb-owt-ucm was shown to be the state of the art on the BSDS, in a large experiment comparing many popular algorithms.

Table 1 shows results of benchmarks on the test set of BSDS, on the validation set of VOC2009 and on the Stanford dataset. The values represent average covering scores of ground truth segmentations by the computed segmentations. *BIS* measures the best covering of the ground truth segmentations by individual segments from any segmentation produced by the evaluated method. *Best* measures for each image the quality of the computed segmentation that best covers the ground truth. *First* evaluates the results using the predicted best segmentation for each image. For gPb-owt-ucm, which does not compute segmentation scores, we select the image scale that produces the best results in the training set, reported as *ODS* in [1]. 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 ground truth segmentations are different, the upper bound for *Best* and *First* 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.

| **BSDS** | *Best* | #T | *First* | *BIS* | #S |
|---|---|---|---|---|---|
| max. possible | .73 | - | .73 | 1.00 | - |
| gPb-owt-ucm | .64 | 64 | **.58** | .74 | 1100 |
| *FG-Tiling* | **.65**/.66 | 64/176 | .57 | **.78** | 194 |

| **VOC2009** | *Best* | #T | *First* | *BIS* | #S |
|---|---|---|---|---|---|
| max. possible | 1.00 | - | 1.00 | 1.00 | - |
| gPb-owt-ucm | .58 | 73 | .45 | .61 | 1043 |
| *FG-Tiling* | **.74**/.76 | 73/140 | **.53** | **.78** | 156 |

| **Stanford** | *Best* | #T | *First* | *BIS* | #S |
|---|---|---|---|---|---|
| max. possible | 1.00 | - | 1.00 | 1.00 | - |
| gPb-owt-ucm | .64 | 58 | .57 | .70 | 503 |
| *FG-Tiling* | **.68**/.70 | 58/180 | **.58** | **.78** | 198 |

Table 1. Average coverings and average number of tilings #T and segments #S computed on the BSDS, VOC2009, and Stanford datasets. The best covering of the ground truth segmentations by any and the predicted best computed segmentation is given by *Best* and *First*, respectively. *Best* is reported both using all tilings computed by *FG-Tiling* and using as many as output by gPb-owt-ucm [1]. *BIS* refers to the best covering of ground truth segmentations by individual segments, not necessarily from a same segmentation. *FG-Tiling* achieves better results than gPb-owt-ucm in all measures except *First* on BSDS.

The results obtained by *FG-Tiling* are superior on both Stanford and VOC2009 and competitive on BSDS. Note that none of the evaluations involve object category recognition. In particular the given VOC2009 scores are *not* using the *semantic segmentation* challenge evaluation which requires recognition. We simply evaluate the quality of unlabeled segmentations (effectively the spatial layout quality of the segments produced) as done previously [1]. The results of gPb-owt-ucm on VOC2009 and Stanford have been computed by us using the code provided by the authors.

We compare different inference procedures for our formulation. Fig. 3 shows the evaluation of *FG-Tiling* and three baselines, *Enum (1min)*, *QP-quadprog*, and *QP-*

*replicator* on the BSDS dataset. 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. 2.

*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 $|\mathcal{S}|$ cliques that have been found[3]. Similar to line 1 of *FG-Tiling, Enum (1min)* first sorts the segments based on $\Phi^t(s_i, \beta_u)$. 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. 3, right. *QP-quadprog* and *QP-replicator* both use the Comtet relaxation [4] but a different solver. *QP-quadprog* uses a generic quadratic program solver, while *QP-replicator* uses the replicator dynamics formulation proposed in [4]. Both solvers return a single local optimum. To obtain multiple tilings we run the solvers several times with different initialization points $\mathbf{y}^0$, one biased towards each vertex $i$: $\mathbf{y}_i^0 = 2/3, \mathbf{y}_j^0 = 1/(3*(|\mathbf{y}|-1))$ for $j \neq i$. *QP-quadprog* produces on average lower quality tilings than *QP-replicator* but also more diverse, which explains the higher coverings for the best over the first $k$ ranked tilings (fig. 3 left).

The quasi-Newton method used in the inner loop of the learning scheme (optimization of feature weights, given a set of computed tilings for the image) required on average less than 300 iterations and completes in less than 15 seconds. The outer loop (uses previously computed weights to compute new tilings and find new weights) 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. 4 shows the progress of learning on the Berkeley Segmentation Dataset (BSDS) [20] (*FG-Tiling* is used for discrete clique optimization).

## 6. Conclusions

We have proposed a mid-level statistical learning and inference framework for image segmentation that given a bag of putative (figure-ground) segmentation hypotheses of an image, selects subsets that form complete interpretations in a principled way. The inference problem is formulated as search for high-scoring maximal weighted cliques in a graph connecting non-overlapping putative figure-ground segment hypotheses. Clique potentials are based on both intrinsic Gestalt segment quality and compatibilities among neighboring image segments, as derived from the projected surface interface statistics of real scenes. We have analyzed empirically the performance of our learning and inference components and have shown that these achieve state-of-the-art segmentation results in the BSDS and Stanford datasets, as well as VOC2009, where the proposed method improves on the state-of-the-art by 31% using the full set of generated tilings, and by 18% for the best predicted tiling. In future work we plan to explore more complex mid-level features, alternative costs for learning, as well as joint models for image segmentation and labeling.

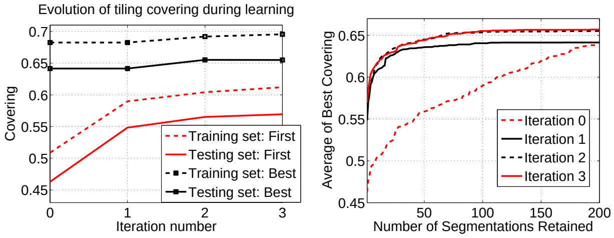

Figure 4. Progress of learning on the BSDS dataset. *Left*: progress of the first ranked and the highest quality tilings on the training and testing sets. Iteration 0 corresponds to the results with the initial weights, iteration 1: the same tilings after the first optimization step, iterations 2–3: after new tilings and learned weights. *Right*: best covering vs. number of segmentations retained on the BSDS test set.

[3] The 1 minute slot given to *Enum (1min)* is about 7.5 × the average run-time of *FG-Tiling* on the BSDS test set. Without the time constraint, *Enum* did not finish enumerating cliques after 48 hours on an image where a pool of $|\mathcal{S}| = 120$ figure-ground segmentations were used.

**Acknowledgements:** This work was supported, in part, by the European Commission, under MCEXT-025481, and by CNCSIS-UEFISCU, under PNII-RU-RC-2/2009.

## References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. *CVPR*, 2009. 1, 2, 5, 6

[2] S. Bagon, O. Boiman, and M. Irani. What is a good image segment? a unified approach to segment extraction. In *ECCV*, 2008. 2

[3] I. Bomze, M. Budinich, P. Pardalos, and M. Pelillo. *Handbook of Combinatorial Optimization*, pages 1–74. Kluwer Academic Publishers, 1999. 3

[4] I. Bomze, M. Pelillo, and V. Stix. Approximating the maximum weight clique using replicator dynamics. *TNN*, 11(6):1228–1241, 2000. 4, 7

[5] W. Brendel and S. Todorovic. Segmentation as maximum-weight independent set. In *NIPS*, pages 307–315, 2010. 2

[6] J. Carreira, A. Ion, and C. Sminchisescu. Image segmentation by discounted cumulative ranking on maximal cliques. *CoRR (arXiv)*, abs/1009.4823, 2010, and TR 06-2010 CVML, University of Bonn. 2

[7] J. Carreira and C. Sminchisescu. Constrained Parametric Min-Cuts for Automatic Object Segmentation. In *CVPR*, 2010. 1, 2, 5

[8] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 24(5):603–619, 2002. 1, 2

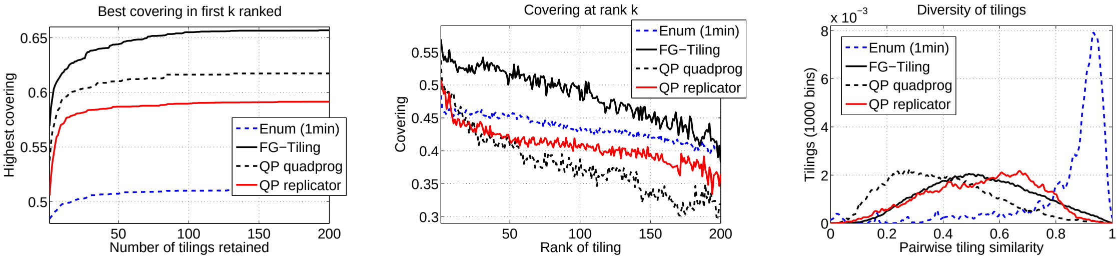

Figure 3. Evaluation of discrete algorithms like *FG-Tiling* and *Enum (1min)* as well as relaxations, *QP-replicator* and *QP-quadprog* on the BSDS dataset. *Left*: best covering for number of tilings considered. *Center*: average covering for given rank (if exists, rank is given by the ordering according to the score in (2)). *Right*: histogram of pairwise similarity in terms of covering measure, between the produced segmentations. See sec. 5 for the definition of the *covering* measure.

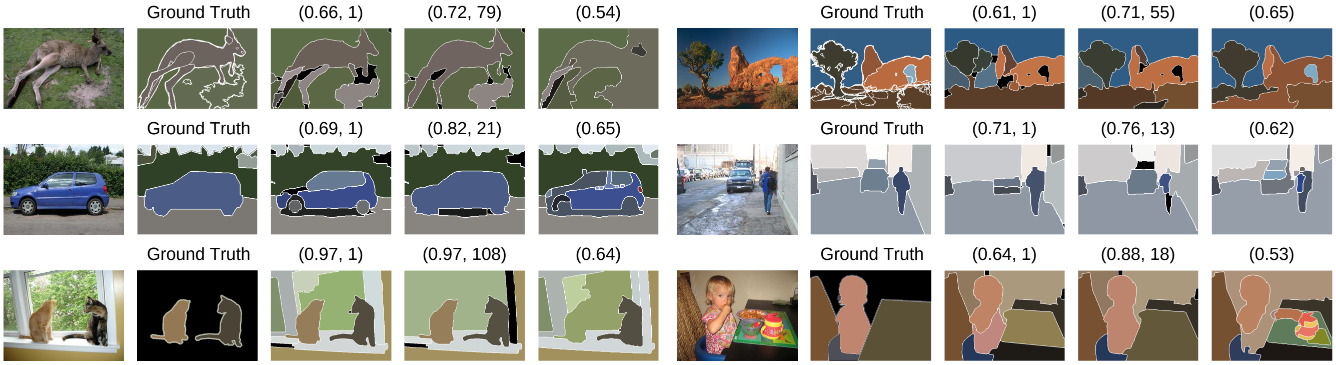

Figure 5. *(Best viewed in color)* Images from the test set of BSDS (first row), Stanford (second row) and VOC2009 (third row). For each image, the second figure is ground truth (GT), the third is our first ranked segmentation (*First* in table 1), whereas the fourth is our closest segmentation to GT (corresponds to *Best* in table 1). The fifth picture is the best segmentation returned by gPb-owt-ucm (also corresponds to *Best*). The multiple ground truths (GT) for BSDS are shown overlaid and the black GT regions in the VOC2009 are regions not annotated. Above segmentations, we show covering scores and the rank produced by our model (2) (for gPb-owt-ucm we show covering). The segmentation ranked first by (2) is usually finer-grained than our closest to GT. Compared to gPb-owt-ucm, the proposed method tends to handle regions with high internal variation better (*e.g.* objects and buildings).

[9] T. Cour, N. Gogin, and J. Shi. Learning spectral graph segmentation. In *AISTATS*, 2005. 2

[10] I. Endres and D. Hoiem. Category independent object proposals. In *ECCV*, pages 575–588, 2010. 2

[11] 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. 5, 6

[12] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004. 1, 2

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

[14] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009. 6

[15] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007. 2

[16] D. Hoiem, A. Stein, A. A. Efros, and M. Hebert. Recovering occlusion boundaries from a single image. In *ICCV*, 2007. 5

[17] P. Huggins, H. Chen, P. Belhumeur, and S. Zucker. Finding folds: On the appearance and identification of occlusion. In *CVPR*, 2001. 5

[18] I. Leichter and M. Lindenbaum. Boundary ownership by lifting to 2.1d. In *ICCV*, 2009. 5

[19] T. Malisiewicz and A. Efros. Improving spatial support for objects via multiple segmentations. In *BMVC*, 2007. 2

[20] 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 *ICCV*, 2001. 6, 7

[21] J. McDermott. Psychophysics with junctions in real images. *JV*, 2(7):131–131, November 2002. 5

[22] X. Ren, C. Fowlkes, and J. Malik. Figure/ground assignment in natural images. In *ECCV*, 2006. 5

[23] X. Ren and J. Malik. Learning a classification model for segmentation. *ICCV*, 2003. 2

[24] E. Sharon, M. Galun, D. Sharon, R. Basri, and A. Brandt. Hierarchy and adaptivity in segmenting visual scenes. *Nature*, 442(7104):719–846, 2006. 2

[25] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000. 1, 2