Implicit scene context for object segmentation and classification

Jan D. Wegner[1], Bodo Rosenhahn[2], and Uwe Soergel[1]

[1]Institute of Photogrammetry and GeoInformation
[2]Institut für Informationsverarbeitung
Leibniz Universität Hannover

Abstract. In this paper, we propose a generic integration of context-knowledge within the unary potentials of Conditional Random Fields (CRF) for object segmentation and classification. Our aim is to learn object-context from the background class of partially labeled images which we call implicit scene context (ISC). A CRF is set up on image super-pixels that are clustered into multiple classes. We then derive context histograms capturing neighborhood relations and integrate them as features into the CRF. Classification experiments with simulated data, eTRIMS building facades, Graz-02 cars, and samples downloaded from Google^{TM} show significant performance improvements.

1 Introduction

Our aim is to segment and classify objects in images. We want to assign a label to each pixel of an image. Context knowledge may add valuable information if local object descriptors deliver ambiguous results in complex scenes. We learn contextual relations between single objects of a scene and introduce them as a prior. Local object descriptors and contextual knowledge are combined in a CRF framework and each pixel is labeled with the most likely object class.

Much research has already focused on how to exploit contextual prior knowledge for object classification in images. In [9] and further related publications Kumar and Hebert extended Conditional Random Fields (CRF), originally proposed in [11], to two-dimensional data and applied them to object detection in images. They consider contextual knowledge through pair-wise potentials that are weighted with features. CRFs provide a highly flexible framework for contextual classification approaches. Torralba et al. [18] use Boosting to learn the graph structure within a CRF framework. Spatial arrangements of objects in an image are learned with a weak classifier and object detection and image segmentation are done in a combined way. Shotton et al. [17] propose an approach based on features derived from texton maps they call "TextonBoost" to achieve joint segmentation and object detection applying Boosting within a CRF framework. Murphy et al. [12] use CRFs for joint object detection and scene classification within a CRF. This classifier learns that particular object categories are more likely to occur in certain scenes than in others. False alarms due to ambiguous local features may be reduced because, for example, polar bears are not likely to appear in a jungle scene. However, this approach considers context on a global scene level but does not model relations of single objects. He et al. [5] introduced the use of a multi-scale CRF for scene segmentation and classification incorporating contextual features at regional and global scene level in addition to local features at pixel-level. Rabinovich et al. [14] formulate a CRF based on image regions that encodes co-occurrence preferences over pair-wise object categories. This allows them to distinguish between object categories that often appear together in the same image and, more important, categories that do usually not appear within the same scene. Calleguillos et al. [3] develop this method further by introducing contextual interactions at pixel-level and at region-level in addition to semantic object interactions via object class co-occurrences. Gould et al. [4] add a spatial component by modelling relative locations between object classes and introducing them into a CRF as additional potential.

Kohli et al. [7] generalize the classical pair-wise Potts model to higher order potentials that enforce label consistency inside image regions. They combine multiple segmentations generated with an unsupervised segmentation method within a CRF for object segmentation and recognition. Related works of Ladicky et al. [10] propose a hierarchical CRF that integrates features computed in different spatial units as pixels, image segments, and groups of segments. They formulate unary potentials over pixels and segments, pair-wise potentials between pixels, and between super-pixels and also a connective potential between pixels and the super-pixels they are contained in.

Heitz and Koller [6] exploit context contained in the background class through what they call the "thing and stuff" (TAS) approach. The main idea is to, first, cluster image super-pixels based both on local features and their ability to serve as context for objects of interest and, second, to integrate this context prior into a rigorous probabilistic framework for object detection. They combine a window detector for local object detection with context that adds predictive power for that particular object category. Savarese et al. [15] compute histograms of so-called correlatons capturing correlations between pairs of pixels based on visual word indices as function of distance. They learn exemplar histograms for each object class from training data and test images are then assigned to the nearest histogram in feature space.

1.1 Contribution

The key idea of our approach is to capture context of the background class of partially labeled images via histograms to support object segmentation and classification. With partially labeled we mean that only a small portion of the object categories existing in the data are semantically annotated in training data. All categories not explicitly labeled are contained within a joint background class. Inspired by the "thing and stuff" (TAS) concept of Heitz and Koller [6] and the "shape context" histograms of Belongie et al. [1] we introduce implicit scene context (ISC) to CRFs. We seek a more general formulation and capture background context and its relation to object classes via histograms (similar to [15]) and integrate it as a potential into a CRF. This is done without major changes to the general CRF framework in terms of training and inference. We do neither add an additional potential nor introduce any complex graph structure but exploit the flexibility provided by the definition of the association potential which depends on all data globally [9].

- Characteristic patterns within the background class of partially labeled images and their relation to labeled object classes are learned.
- Contextual patterns are formulated in terms of histograms. We achieve rotation invariance and the use of multiple context scales ensures good performance for both small and big objects.
- Although we model it as a unary potential within a CRF framework it can generally be utilized (with minor changes) with any kind of non-contextual classifier like Support Vector Machines, too.

This novel approach is generally applicable to any kind of image scene, for example, aerial images, terrestrial images, and medical images.

2 CRF classification framework

In the following, we denote scalars in normal face type and vectors in bold face type. CRFs are discriminative models and thus directly model the posterior distribution $P\left(\mathbf{y}|\mathbf{x}\right)$ of the labels $\mathbf{y}$ given data $\mathbf{x}$. The label of the node i of interest is y_i and y_j the label of node j it is compared to. We have to formulate a cost function which is usually written as an energy term $E\left(\mathbf{x},\mathbf{y}\right)$ that encapsulates unary potentials and pair-wise potentials. In order to gain a posterior distribution $P\left(\mathbf{y}|\mathbf{x}\right)$ we need to turn the energies into probabilities by normalizing them through the partition function $Z\left(\mathbf{x}\right)$. Making use of sufficient statistics of the exponential family we may then write the posterior distribution $P\left(\mathbf{y}|\mathbf{x}\right)$ as:

$$P\left(\mathbf{y}|\mathbf{x}\right) = \frac{1}{Z\left(\mathbf{x}\right)} \exp\left(E\left(\mathbf{x},\mathbf{y}\right)\right) \tag{1}$$

Following the notations of Kumar and Hebert [9] we can express the energy term $E\left(\mathbf{x},\mathbf{y}\right)$ as the sum of association potentials $\mathbf{A}_i\left(\mathbf{x},y_i\right)$ and interaction potentials $\mathbf{I}_{ij}\left(\mathbf{x},y_i,y_j\right)$:

$$E\left(\mathbf{x},\mathbf{y}\right) = \sum_{i\in S} A_i\left(\mathbf{x},y_i\right) + \sum_{i\in S}\sum_{j\in N_i} I_{ij}\left(\mathbf{x},y_i,y_j\right) \tag{2}$$

The association potential $\mathbf{A}_i\left(\mathbf{x},y_i\right)$ measures how likely a label site i is labeled with y_i given the data $\mathbf{x}$. It contains all unary potentials defined over cliques of size one and this is where our implicit context will be incorporated. The interaction potential $\mathbf{I}_{ij}\left(\mathbf{x},y_i,y_j\right)$ models the pair-wise potentials that are defined over cliques of size two. It describes how two label sites i and j interact and we will leave this term almost unchanged.

Both potentials, unary and pair-wise, have access to all data $\mathbf{x}$ of the set S of all image sites. Additionally, the pair-wise potentials also have access to all labels $\mathbf{y}$ globally because the neighborhood N_i of site i of $\mathbf{I}_{ij}(\mathbf{x}, y_i, y_j)$ may potentially be the entire image. Those properties of CRFs provide a high degree of flexibility and we can thus introduce context from very local to global scales into both terms of the energy term in Eq. 2. However, the standard modelling of the association potential $\mathbf{A}_i(\mathbf{x}, y_i)$ and the interaction potential $\mathbf{I}_{ij}(\mathbf{x}, y_i, y_j)$ (e.g., [9]) does not fully exploit the possibility of considering labels $\mathbf{y}$ and given data $\mathbf{x}$ globally. Much research effort has gone into finding a more general and global formulation of context through label comparisons in the interaction potential (e.g., [14, 7, 3]). Our focus is on exploiting the full flexibility provided by the CRF definition of the unary potentials of $\mathbf{A}_i(\mathbf{x}, y_i)$. We seek a more general and global incorporation of all data $\mathbf{x}$ as done, for example, by Murphy et al. [12]. If we model the association potential $\mathbf{A}_i(\mathbf{x}, y_i)$ as a linear model the standard formulation is:

$$A_i(\mathbf{x}, y_i) = y_i \mathbf{w}^T \mathbf{h}_i(\mathbf{x}) . \tag{3}$$

Node features $h_i(\mathbf{x})$ generated from data $\mathbf{x}$ are contained in vector $\mathbf{h}_i(\mathbf{x})$ and the corresponding weights, which are tuned during the training process, are contained in vector $\mathbf{w}^T$. We will integrate ISC through the feature vector and thus $\mathbf{h}_i(\mathbf{x})$ will be replaced as we will explain in section 3.1. The interaction potential $\mathbf{I}_{ij}(\mathbf{x}, y_i, y_j)$ determines how two sites i and j should interact regarding all data $\mathbf{x}$ (see Eq. 4). Using again a linear model we can write:

$$I_{ij}(\mathbf{x}, y_i, y_j) = y_i y_j \mathbf{v}^T \boldsymbol{\mu}_{ij}(\mathbf{x}) . \tag{4}$$

$\boldsymbol{\mu}_{ij}(\mathbf{x})$ contains all edge features and $\mathbf{v}^T$ the weights, respectively. Edge features $\boldsymbol{\mu}_{ij}(\mathbf{x})$ can generally be chosen based on any kind of feature derived from data $\mathbf{x}$. They should however somehow reflect and model the relationship of the nodes i and j that are compared. The standard approaches consist of either concatenating the feature vectors $\mathbf{h}_i(\mathbf{x})$ and $\mathbf{h}_j(\mathbf{x})$ of both nodes or of subtracting them element-wise. We choose the latter one and $\boldsymbol{\mu}_{ij}(\mathbf{x})$ is:

$$\boldsymbol{\mu}_{ij}(\mathbf{x}) = |\mathbf{h}_i(\mathbf{x}) - \mathbf{h}_j(\mathbf{x})|. \tag{5}$$

3 Implicit scene context (ISC)

The idea is to exploit spatial patterns contained in the background class of a partially labeled image to support object segmentation and classification. We can then benefit from very large image databases where images are only partially labeled and learn context although we do not explicitly know all object classes. In addition to the object classes that have been explicitly labeled for training we can use patterns existing in the unlabeled part of the data (i.e., labeled as background class).

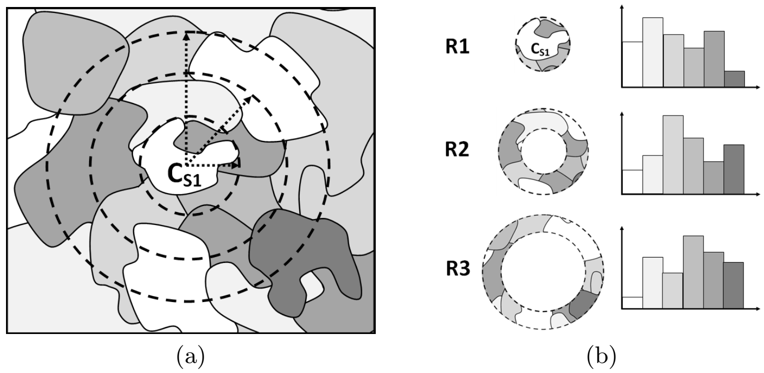


Fig. 1. Principle of implicit context: (a) ranges around the centroid C_{S1} of image super-pixel $S1$ (grey levels indicate different labels appointed to the super-pixels with k-means during training or NN during testing), (b) histograms of cluster labels of the three ranges $R1$, $R2$ and $R3$.

The following requirements have to be met: We should be able to cope with very local to global context scales. In addition, we want to keep ISC generically applicable to multiple kinds of scenes. For example, it should capture context in terrestrial images of building facades where usually sky is above the facade and vegetation below but also in aerial images of buildings where no preferred ordering with attributes like "above" and "below" exists. Thus, we do not want to rely on any kind of preferred direction. Finally, we want to achieve computational efficiency and avoid the computation of co-occurrences. In order to meet these requirements we take the following steps that will be explained in detail in the following paragraphs:

- multi-scale image segmentation into super-pixels and feature computation,
- unsupervised k-means clustering and nearest-neighbor (NN) classification of the super-pixels based on the previously generated features,
- generation of context histograms in three different ranges per super-pixel,
- input as feature vector to the CRF unary potentials.

3.1 Context potential within CRF

During training we first perform an unsupervised classification of all super-pixels. We could use any kind of unsupervised classifier but for means of speed and simplicity we chose a k-means clustering followed by a NN classification. As input to the k-means clustering we use all features $\mathbf{h}_i(\mathbf{x}) \in \mathbf{h}(\mathbf{x})$ that were computed per super-pixel. The exact cluster centers $\mathbf{K}$ we compute with the k-means clustering $\mathbf{K} = K_{means}(\mathbf{h}(\mathbf{x}))$ are used for the following processing.

Each super-pixel is labeled with $y_{us,i} \in \mathbf{y}_{us}$ where $\mathbf{y}_{us}$ contains all unsupervised labels corresponding to the number of chosen cluster centers k. Label

$y_{us,i}$ of the super-pixel i of interest is determined via NN and is thus a function of the minimum mean distance between feature vector $h_i(\mathbf{x})$ and the cluster centers $\mathbf{K}$. Each super-pixel i is assigned the cluster center $\mathbf{K}_c$ (where $c = 1...k$ is the cluster center with k the total number of all cluster centers) that is the closest in feature space. The resulting labeled super-pixels (e.g., with $k = 6$) are shown schematically in Fig. 1(a). Next, the centroid C_S of each super-pixel is determined and histograms of labels $hist_R(\mathbf{y}_{us})$ occurring within three different ranges R around each super-pixel are generated. The number of label occurrences $\mathbf{y}_{us}$ within each range R is counted (Fig. 1(b) with three ranges $R1$, $R2$, and $R3$). We can choose either short or long ranges depending on whether we would like to incorporate local or global context, respectively. It should be noted that longer ranges do not lead to any more complex graph structure because no graph is set up at this point. Furthermore, the number of the ranges and either coarse or fine scaling enables us to capture the distribution of object categories contained in the background class as a function of their distance to the node of interest. Then, various moments and additional information representing the contextual patterns in the environment of a particular super-pixel are derived from the histograms. We use qualitative, quantitative, and spatial context features $\mathbf{C}(\mathbf{h}(\mathbf{x}))$ (e.g., most often occurring label). For testing we apply exactly the same processing steps but drop the k-means clustering. The cluster centers $\mathbf{K}$ that were determined during training are passed to testing and the NN cluster centers to the nodes of the test data are computed. Thus, all super-pixels of the test data are labeled corresponding to the unsupervised classification performed during training. The implicit context features $\mathbf{C}_i(\mathbf{h}(\mathbf{x}))$ of the test data are computed and introduced into a linear model:

$$A_i(\mathbf{x}, y_i) = y_i \mathbf{w}^T \mathbf{C}_i(\mathbf{h}(\mathbf{x})) \tag{6}$$

We can then either determine the class of each super-pixel i merely based on implicit context features $\mathbf{C}_i(\mathbf{h}(\mathbf{x}))$ or also add the local node features $\mathbf{h}_i(\mathbf{x})$ to the feature vector. The pair-wise potentials only change in such a way (cf. Eq. 5) that the element-wise absolute differences between nodes i and j in the graph are now computed based on the corresponding implicit context features:

$$\boldsymbol{\mu}_{ij}(\mathbf{x}) = abs(\mathbf{C}_i(\mathbf{h}(\mathbf{x})) - \mathbf{C}_j(\mathbf{h}(\mathbf{x}))) \tag{7}$$

We do not perform any normalization of the label count in the histogram, for example, based on the size of the super-pixels because tests show that the importance of a super-pixel does not necessarily increase with its size. In other words, small super-pixels may be characteristic context features and thus are of high relevance for a particular object class.

4 Experiments

We perform several experiments with partially labeled data in order to assess the benefits of ISC-CRF. Only one object category is semantically annotated

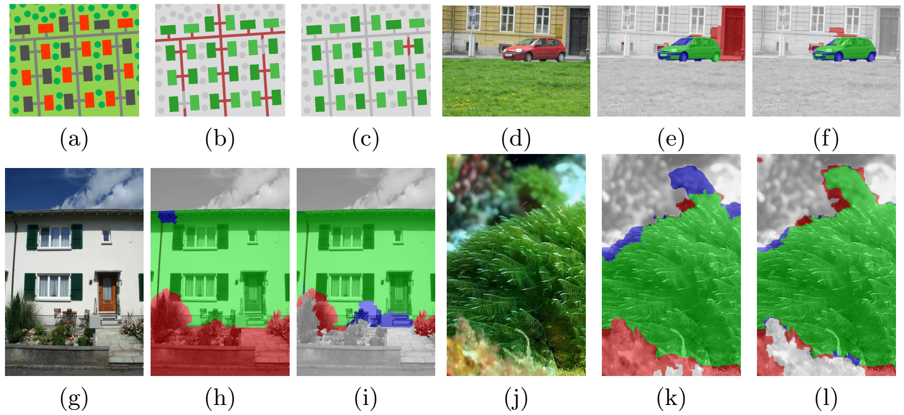


Fig. 2. Results with simulated data, eTRIMS [8] facade images, algae, and Graz-02 cars [13]: true positives (green), false positives (red) and false negatives (blue) without implicit scene context (b, e, h, k) and with implicit scene context (c, f, i, l).

in training data and all other categories are labeled as background. First, we demonstrate the performance improvements achieved with ISC-CRF compared to a standard CRF for different object classes and background patterns (4.1). Second, we evaluate the impact of different cluster center numbers and, third, we assess the robustness to noise (4.2). Quickshift [19] is used for super-pixel generation. If a super-pixel extends across an object boundary it may not be repaired later on in the process. We thus over-segment all images to ensure consistency of object boundaries and super-pixels. In order to avoid unstable feature distributions of too small super-pixels we generate a segmentation in three different scales. Super-pixels sharing a common boundary at the highest scale are linked with edges in the graph. Features of coarser-scale super-pixels are written to the vectors of the highest-scale super-pixels they contain. As features $\mathbf{h}_i(\mathbf{x})$ of a super-pixel we compute the first two moments of the color information and oriented gradient histogram features. We select those very simple features for reasons of transparency and ease of replicability. A subset of different benchmark data sets (nine images out of each) is used to verify the proposed ISC-CRF concept. A quadratic expansion of the feature vectors is done as described by Kumar and Hebert [9] in order to introduce a more precise quadratic decision surface. We apply the quasi-Newton method limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) for parameter estimation and loopy belief propagation for approximate inference using Mark Schmidt's toolbox [16]. Cross-validation is performed with two thirds of the data for training and one third for testing (as recommended by Crowther and Cox [2]) in order to compute true positive rate (TPR) and false positive rate (FPR) pixel-wise. The TPR is the percentage of all correctly labeled object pixels and the FPR is the percentage of all background pixels that are misclassified as object.

Table 1. TPR and FPR in % of standard CRF and ISC-CRF

	CRF		ISC-CRF	
Data	TPR	FPR	TPR	FPR
Simulation	85.9	**6.8**	85.9	**0.8**
eTRIMS facades	86.9	**22.1**	88.1	**7.3**
Algae	75.7	**37.0**	84.5	**23.7**
Graz-02 cars	86.6	**16.4**	88.1	**4.3**

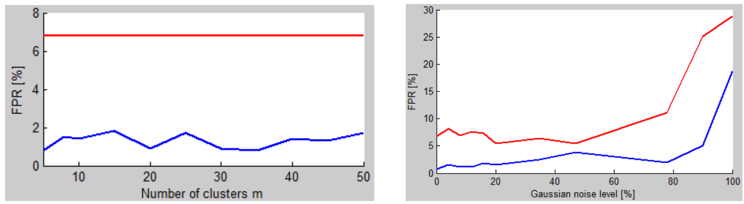
4.1 Classification of objects in different scenes

In order to verify the general applicability of the implicit scene context we perform tests with four different object class scenes: with simulated aerial images of an urban scene, with facade images taken from the eTRIMS benchmark data [8], with Google^{TM} images of algae and with car images of the Graz-02 benchmark data [13]. Those four object class categories are chosen because they represent different spatial object and background distributions. Many small objects (buildings) embedded into background context are contained in the simulated urban scene (Fig. 2(a)). Small irregular objects entirely surrounded by background context are the cars (Fig. 2(d)), single very large objects (facades) with clear straight boundaries and with background context only above and below are the building facades (Fig. 2(g)) and large but frayed objects partially surrounded by background context are the algae (Fig. 2(j)). A good performance of the implicit scene context approach for all tasks would support the claim of general applicability to any kind of image scene.

The classification performance results of the test data are summarized in Table 1. Example images and the corresponding results are shown in Fig. 2. In all four cases the ISC-CRF decreases the FPR significantly in comparison to the standard CRF. On an Intel^{TM} Core i7 2.4 Ghz CPU, 12 GB RAM the computation time using the implicit scene context potential does only marginally increase by several seconds per image.

4.2 Parameter assessment with simulated data

The context ranges, the number of k-means cluster centers, and the segmentation scales are currently adapted manually to each data set. The previously introduced simulated urban scene (see example in Fig. 2(a)) is used to evaluate the impact of varying cluster centers because we know the exact number of object categories contained in the data: buildings (red and gray rectangles), trees (dark green circles), grassland (light green background) and streets (light gray lines). The buildings are our labeled object class and all other object categories are contained in the background class. Only color features are used for these tests leading to five distinct clusters due to the building class consisting of red and dark gray buildings. We use three different ranges (10, 20, and 30 pixel radii) and perform tests with five up to 50 cluster centers. The FPR of each test is



(a) (b)

Fig. 3. FPR of ISC-CRF (blue) and standard CRF (red) classification of simulated data: (a) with varying numbers of k-means cluster centers and (b) with different noise levels.

displayed in blue Fig. 3(a) whereas the FPR of the standard CRF is displayed in red. The FPR varies about 1 % (from 0.8 % to 1.8 %) and no significant trend is observable. Changing the number of k-means cluster centers has a very small impact on the classification performance but of course on computation time. A rather small number of cluster centers is beneficial. The radii of the context ranges and the segmentation scale are adapted to each scene separately because both parameters depend on the scales of context and objects. This makes the ISC-CRF highly flexible and easy to adapt to new scenes. Both parameters could also be introduced into the learning step without major changes to the general framework.

Second, we test if the ISC-CRF is robust to image noise and whether we gain robustness compared to a standard CRF. Several gaussian noise levels with mean zero and standard deviations up to 100 % (corresponding to 256 in our case of 8 bit RGB channels) are generated and added to the RGB channels of the simulated data, which is then cropped in order to keep all values between zero and 255. Cross-validation tests with CRF and ISC-CRF is done and FPR is recorded. In figure 3(b) FPR of the standard CRF (red) and FPR of ISC-CRF (blue) of all tested noise levels are displayed. For all noise levels the FPR of the ISC-CRF stays below that of the standard CRF. Furthermore, the ISC-CRF is slightly more robust to noise because its FPR starts increasing later (approx. 90 % vs. approx. 80 %).

5 Conclusions and future work

In this paper we have introduced the concept of implicit scene context to learn context in an unsupervised way from the background class. Tests with four different scene types have shown that the ISC-CRF decreases the FPR while increasing the TPR compared to a standard CRF. We have demonstrated that different spatial object and background distributions can be captured via the context histograms. In future work we want to integrate more complex features and feature combinations, test our method on complete benchmark datasets, and learn those parameters that are currently chosen empirically.

References

1. Belongie, S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(24), 509–522 (2002)
2. Crowther, P.S., Cox, R.J.: A Method for Optimal Division of Data Sets for Use in Neural Networks. In: et al., R.K. (ed.) LNCS, pp. 1–7. Springer-Verlag, Berlin Heidelberg (2005)
3. Galleguillos, C., McFee, B., Belongie, S., Lanckriet, G.: Multi-Class Object Localization by Combining Local Contextual Interactions. In: CVPR (2010)
4. Gould, S., Rodgers, J., Cohen, D., Elidan, G., Koller, D.: Multi-Class Segmentation with Relative Location Prior. International Journal of Computer Vision 80(3), 300–316 (2008)
5. He, X., Zemel, R.S., Carreira-Perpiñán, M.: Multiscale Conditional Random Fields for Image Labeling. In: CVPR (2004)
6. Heitz, G., Koller, D.: Learning Spatial Context: Using Stuff to Find Things. In: ECCV (2008)
7. Kohli, P., Ladicky, L., Torr, P.H.: Robust Higher Order Potentials for Enforcing Label Consistency. International Journal of Computer Vision 82(3), 302–324 (2009)
8. Korč, F., Förstner, W.: eTRIMS Image Database for interpreting images of man-made scenes. Tech. Rep. TR-IGG-P-2009-01 (April 2009)
9. Kumar, S., Hebert, M.: Discriminative Random Fields. International Journal of Computer Vision 68(2), 179–201 (2006)
10. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.: Associative Hierarchical CRFs for Object Class Image Segmentation. In: ICCV (2009)
11. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for segmenting and labeling sequence data. In: ICML (2001)
12. Murphy, K.P., Torralba, A., Freeman, W.T.: Using the Forest to See the Trees: A Graphical Model Relating Features, Objects, and Scenes. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA (2004)
13. Opelt, A., Pinz, A., Fussenegger, M., Auer, P.: Generic Object Recognition with Boosting. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(3), 416–431 (2006)
14. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in Context. In: ICCV (2007)
15. Savarese, S., Winn, J., Criminisi, A.: Discriminative Object Class Models of Appearance and Shape by Correlatons. In: CVPR (2006)
16. Schmidt, M.: UGM: A Matlab toolbox for probabilistic undirected graphical models, http://www.cs.ubc.ca/ schmidtm/Software/UGM.html accessed 15th July 2010
17. Shotton, J., Winn, J., Rother, C., Criminisi, A.: TextonBoost: Joint Appearance, Shape and context modeling for Multi-class Object Recognition and Segmentation. In: ECCV (2006)
18. Torralba, A., Murphy, K.P., Freeman, W.T.: Contextual Models for Object Detection Using Boosted Random Fields. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems, pp. 1401–1408. MIT Press, Cambridge, MA (2005)
19. Vedaldi, A., Soatto, S.: Quick shift and kernel methods for mode seeking. In: ECCV (2008)