DISCRIMINATIVE DICTIONARY LEARNING WITH SPATIAL PRIORS

Nazar Khan, Marshall F. Tappen

University of Central Florida
School of Electrical Engineering and Computer Science, Orlando, FL
nazar,mtappen@cs.ucf.edu

ABSTRACT

While smoothness priors are ubiquitous in analysis of visual information, dictionary learning for image analysis has traditionally relied on local evidences only. We present a novel approach to discriminative dictionary learning with neighborhood constraints. This is achieved by embedding dictionaries in a Conditional Random Field (CRF) and imposing labeldependent smoothness constraints on the resulting sparse codes at adjacent sites. This way, a smoothness prior is used while learning the dictionaries and not just to make inference. This is in contrast with competing approaches that learn dictionaries without such a prior. Pixel-level classification results on the Graz02 bikes dataset demonstrate that dictionaries learned in our discriminative setting with neighborhood smoothness constraints can equal the state-of-the-art performance of bottom-up (i.e. superpixel-based) segmentation approaches.

Furthermore, we isolate the benets of our learning formulation and CRF inference to show that our dictionaries are more discriminative than dictionaries learned without such constraints even without CRF inference. An additional benet of our smoothness constraints is more stable learning which is a known problem for discriminative dictionaries.

Index Terms— Dictionary Learning, Smoothness Prior, Pixel-level Classicaiton, Segmentation, Discriminative

1. INTRODUCTION

Discriminative learning of sparse-coding based dictionaries has been shown to improve performance on various computer vision tasks. Interestingly, while these dictionaries are often eventually used for analyzing natural images which are characterized by a local smoothness prior, no such local neighborhood context is used in the dictionary learning process. We show how to discriminatively learn dictionaries while enforcing smoothness constraints from the local spatial neighborhoods. This is done by embedding the dictionary learning framework in a Conditional Random Field (CRF).

Dictionary learning has successfully been used for various signal classication tasks such as pixel-level classication of images [1, 2, 3, 4], object localization [5], image classication [6], face recognition [7] and video classication [8, 9].

M.F.T. was supported by NSF grants IIS-0905387 and IIS-0916868.

	[11]	[12]	Ours
Structured Prediction	×	✓	✓
Smoothness Constraints	✓	×	✓
Per-class Dictionaries	×	×	✓
Linear Classifier	✓	✓	×

Table 1. Comparison with closely related approaches.

Standard approaches learn dictionaries either reconstructively [10] or discriminatively [1, 2, 3, 6, 4] but do not attempt to exploit neighborhood context in the learning process.

Images of real world objects in real world settings exhibit strongly smooth labels. Generally, pixels from a certain class lie adjacent to each other. This calls for a smoothness prior in the energy formulation and it allows us to enforce smoothness constraints on neighboring sparse code pairs for a dictionary. But since boundaries of objects do not share this smoothness prior, there is a need for a discontinuity preserving prior too. This discontinuity preserving prior is what allows us to enforce (non-)smoothness constraints between dictionariesfrom different classes. *To the best of our knowledge, this is the rst attempt at learning discriminative dictionaries for pixel classication with label and location dependent sparse code (non-)smoothness constraints*.

The closest related work in terms of smoothness constraints is that of Guo *et al.* [11] which uses sparse code smoothness constraints for image classication. The key difference from their work is that we operate on the pixel level and therefore ours is a structured prediction problem while theirs is a standard classication problem. For a given image, they infer a single label while we infer the pixel labelling structure. Another closely related work is that of Yang & Yang [12] that also embeds a dictionary in a CRF. However, despite making use of the structured CRF grid, they impose no smoothness constraints during dictionary learning. Mairal *et al.* [13] use simultaneous sparse coding whereby similar image patches are encouraged to have similar sparse codes. We use the same intuition but for learning dictionaries instead of sparse code computation and we use a neighborhood structure instead of patch similarity. Table 1 summarizes the relationships between our work and its closest counterparts.

Besides increased discriminability, an additional benefit of our smoothness constraints is the mitigation of numerical instability which is inherent to discriminative dictionary

learning [1, 4]. Interestingly, a recent stability analysis [14] for reconstructive dictionaries also concluded that sparse code smoothness plays an important role in stable learning.

2. PRELIMINARIES

For an image $\mathbf{y}$ with ground-truth labeling $\mathbf{x}$, let $\mathcal{V}$ be a uniformly spaced grid of image locations or 'sites'and $\mathbf{y}_i \in \mathbb{R}^n$ be an n dimensional feature vector extracted at site $i \in \mathcal{V}$. For each site i, $\mathcal{N}_i$ denotes the neighboring sites of i and $x_i \in \{1 \dots C\}$ denotes the true label.

For each feature vector $\mathbf{y}_i \in \mathbb{R}^n$, let $\mathbf{s}_{ic} \in \mathbb{R}^k$ be its sparse code vector under a dictionary $\mathbf{D}_c \in \mathbb{R}^{n \times k}$ for class $c \in \{1 \dots C\}$. The sparse code vector $\mathbf{s}_{ic}$ is obtained as a solution to the ℓ_1 sparse coding problem

$$\mathbf{s}_{ic}(\mathbf{y}_i, \mathbf{D}_c) = \arg\min_{\mathbf{s} \in \mathbb{R}^k} \frac{1}{2}||\mathbf{y}_i - \mathbf{D}_c\mathbf{s}||_F^2 + \lambda||\mathbf{s}||_1 \quad (1)$$

The reconstruction error $\mathbf{R}_{ic}$ for a signal $\mathbf{y}_i$ under a dictionary $\mathbf{D}_c$ is computed using the optimal sparse code vector $\mathbf{s}_{ic}$ obtained via (1)

$$\mathbf{R}_{ic}(\mathbf{y}_i, \mathbf{D}_c) = \frac{1}{2}||\mathbf{y}_i - \mathbf{D}_c\mathbf{s}_{ic}||_F^2 + \lambda||\mathbf{s}_{ic}||_1 \quad (2)$$

$\mathbf{R}_i \in \mathbb{R}^C$ denotes the vector of per-class reconstruction errors for signal $\mathbf{y}_i$. Both (1) and (2) are rendered non-differentiable with respect to dictionary $\mathbf{D}_c$ due to the presence of the ℓ_1 norm. However, implicit differentiation can be employed to compute such gradients. Details can be found in [15, 16, 12].

3. DISCRIMINATIVE DICTIONARY LEARNING WITH SPATIAL PRIOR

Let $\mathbf{Y} = [\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(N)}]$ be N training images with corresponding labelings $\mathbf{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}]$. Let $\mathbf{S}_i = [\mathbf{s}_{i1}, \dots, \mathbf{s}_{iC}]$ denote the matrix of sparse codes of signal $\mathbf{y}_i$ under each of the C dictionaries. Without loss of generality, let L be the set of all possible labelings on any given grid of sites. Clearly, L is an exponentially large set. Then the probability of image labeling $\mathbf{x}^{(t)}$ conditioned on the observed image $\mathbf{y}^{(t)}$ can be written as a Gibbs field

$$P(\mathbf{x}^{(t)}|\mathbf{y}^{(t)}, \{\mathbf{D}\}_1^C, \boldsymbol{\kappa}) = \frac{1}{Z} e^{-E(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \{\mathbf{D}\}_1^C, \boldsymbol{\kappa})} \quad (3)$$

where $Z = \sum_{\mathbf{x} \in \mathrm{L}} e^{-E(\mathbf{x}, \mathbf{y}^{(t)}, \{\mathbf{D}\}_1^C, \boldsymbol{\kappa})}$ is the so-called partition function and

$$\begin{aligned} E(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \{\mathbf{D}\}_1^C, \boldsymbol{\kappa}) &= \sum_{i \in \mathcal{V}^{(t)}} E_i(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \{\mathbf{D}\}_1^C, \boldsymbol{\kappa}) \\ = \sum_{i \in \mathcal{V}^{(t)}} & e^{-\kappa^{\text{data}}} \underbrace{\left(\mathcal{D}_{x_i}^{\mathbf{R}_i} + e^{-\kappa^{\text{rec}}} \mathbf{R}_{ix_i}\right)}_{\text{data term}} \\ & + e^{-\kappa^{\text{smooth}}} \underbrace{\sum_{j \in \mathcal{N}_i} e^{-\kappa^{\text{ind}}} \bar{\delta}_{x_i x_j} + e^{-\kappa^{\text{dep}}} s_{\text{dep}}}_{\text{smoothness term}} \end{aligned} \quad (4)$$

where $\mathcal{D}_{x_i}^{\mathbf{R}_i} = \mathbf{R}_{ix_i} \bar{\mathbf{R}}_i$ is the discriminative deviation function [4] that encourages the reconstruction error for the true class x_i to be lowest among all classes. This leads to greater discriminability. Value of κ^{rec} determines the weightage given to the reconstructive term relative to the discriminative deviation term. The data-independent smoothness term $\bar{\delta}_{x_i x_j}$ penalizes dissimilar labels on adjacent sites and rewards similar labels. The data-dependent smoothness term is

$$s_{\text{dep}} = -\delta_{x_i x_j} \left(\mathcal{D}_{x_i}^{\mathbf{p}} + \mu \mathbf{p}_{x_i}\right) + \bar{\delta}_{x_i x_j} \mathbf{p}_{x_i} \quad (5)$$

where

$$\mathbf{p} = \mathbf{s}_{ix_i}^T \mathbf{s}_j \quad (6)$$

is a C dimensional vector of the similarity of site is sparse code under dictionary $\mathbf{D}_{x_i}$ with all sparse codes of the adjacent site j. Entry $\mathbf{p}_k$ denotes the similarity of sparse codes $\mathbf{s}_{ix_i}$ and $\mathbf{s}_{jx_k}$. Here too, discriminative deviation is employed to encourage smoothness of sparse codes at adjacent sites that belong to the same class. The weights of the data term and the smoothness term are determined by parameters κ^{data} and κ^{smooth} respectively. Negative exponentials of all weights are used to ensure positive weightings and unconstrained optimization.

Data-dependent smoothness. The goal is to encourage signals with the same label to have similar sparse code vectors and those with different labels to have dissimilar sparse code vectors. During learning, this encourages dictionaries to be more sensitive to object boundaries. During inference, this allows smoothing to be reduced at edges (in feature space) and results in sharper segmentations. For adjacent pixels i, j with the same label $x_i = x_j$, the data-dependent smoothness term encourages sparse code vectors $\mathbf{s}_{ix_i}$ and $\mathbf{s}_{jx_i}$ under dictionary $\mathbf{D}_{x_i}$ to be most similar among all classes. This is achieved by once again employing the discriminative deviation function as used in the data term. *The advantage of using discriminative deviation is dictionary learning with label-dependent smoothness constraints on adjacent sparse codes*. If only the term $\mathbf{s}_{ix_i}^T \mathbf{s}_{jx_i}$ is used instead, then only dictionary $\mathbf{D}_{x_i}$ is affected. Parameter $\mu \geq 0$ determines the trade-off between discriminative deviation and the similarity of the sparse code vectors. For adjacent pixels with different labels, the sparse code vectors only under dictionary $\mathbf{D}_{x_i}$ are encouraged to be different. Since our graphical model contains loops, this eventually implies sparse code dissimilarity under both classes x_i and x_j. However, no inter-dictionary constraint is enforced in this case.

Energy function (4) makes our formulation a Discriminative Random Field (DRF) [17] which is a variant of a Conditional Random Field (CRF) [18]. Instead of learning linear CRF parameter vectors, we learn non-linear dictionaries. It has similarities with [12] but has a richer representational model and includes data-dependent smoothness. Our formulation tries to explain all classes instead of just foreground. *More importantly, the data-dependent smoothness term includes, in addition to the data, the dictionaries as well. During learning, this encourages dictionaries to have responses*

for neighboring pixels that reflect their labels. Therefore, energy function (4) *imposes neighborhood constraints on the discriminative dictionary learning frameworks from [1, 4]. It can also be viewed as the structured prediction counterpart of [11].*

Stability. Our smoothness constraints can alternatively be considered as *pseudo-regularization* of dictionaries based on the regularity of pixel labels in natural images. It is well-known that

1. Sparse coding is sensitive to incoherence among a dictionary's atoms [10], and

2. Discriminability is increased by having mutually incoherent dictionaries [3].

Therefore, it is beneficial to increase both intra- and inter-dictionary incoherence. Intra-dictionary incoherence is enforced by $\bar{\delta}_{x_i x_j}\mathbf{p}_{x_i}$ in Equation (5). The discriminative deviation term $\mathcal{D}^{\mathrm{p}}_{x_i}$ enforces inter-dictionary incoherence and also leads to well-conditioned dictionaries by requiring adjacent same-class sparse codes to be similar. One indicator of ill-conditioned dictionaries is that the sparse coding yileds very large values [14]. So conversely, by requiring adjacent sparse codes to be (typically) similar, the dictionaries are encouraged to be well-conditioned. So our formulation contains the well-known sources of stability and discrimination. In contrast, despite embedding dictionary learning in a CRF framework, Yang & Yang [12] do not impose such dictionary-related smoothness constraints.

Inference and Parameter Learning. Computation of (3) and its corresponding likelihood function require computation of the partition function Z which is intractable due to the exponentially large size of the set L of all labelings. Therefore, for inference we use approximate techniques such as Mean Field Inference or Loopy Belief Propagation. For learning parameters, we use the pseudolikelihood approximation which replaces the intractable computation of the true partition function by the tractable computations of local normalization functions. A potential disadvantage of pseudolikelihood is over-smoothed MAP inference [19] and this is handled by learning the weights in $\boldsymbol{\kappa}$.

Initialization. $\mathbf{D}$: Dictionaries can be initialized to be random or obtained through K-means, K-SVD or any other reconstructive or even discriminative dictionary learning technique. In order to allow a fairer comparison with [12], we initialize via K-means.

$\boldsymbol{\kappa}$: Inference on the random field in (4) is very sensitive[1] to the smoothness weights κ^{smooth}, κ^{dep}, and κ^{ind}. Therefore, before learning, it is important to properly initialize them. Initializing $\boldsymbol{\kappa} = \{\kappa^{\mathrm{data}}, \kappa^{\mathrm{rec}}, \kappa^{\mathrm{smooth}}, \kappa^{\mathrm{ind}}, \kappa^{\mathrm{dep}}\}$ to $\{-2, -3, -1, 3, 10\}$ was emiprically found to be a good starting point.

[1] [12], for instance, do not attempt to learn their smoothness weight $\mathbf{w}_2$ for this reason.

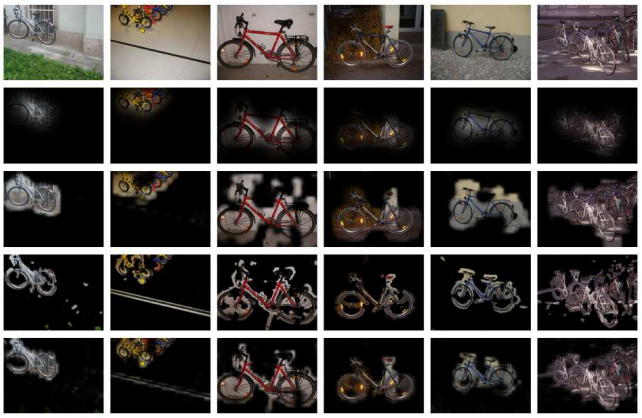

Fig. 1. Pixel-wise classification results for some test images from the Graz02 bike dataset. **1st Row:** Original. **2nd Row:** Khan & Tappen [4] with vanilla Gaussian smoothing on raw classification. **3rd Row:** Yang & Yang [12] (CRF with Potts model). The advantages of using boundary-preserving smoothness can be clearly observed in **4th Row:** Our CRF inference on a grid with spacing of 4 pixels followed by interpolation. **5th Row:** Our CRF inference with classification on a grid with spacing of 20 pixels followed by interpolation. The labellings of [4] and [12] appear to be over-smoothed and can tend to cross over object boundaries. While visually inferior, such over-smoothing can lead to inated quantitative results as hinted in [20]. Implementation of [12] was made available by the original authors.

CRF+Dictionary		Dictionary		Shape Mask
Ours	[12]	[4]	[3]	[21]
72.1	62.4	69.5	68	61.8

Table 2. Comparison of EER (%) of precision-recall curves for pixel-level classfication of Graz02 bike test set. Our results exceed the state-of-the-art in top-down dictionary learning based approaches and match the bottom-up super-pixel based segmentation approach in [22].

4. EXPERIMENTS AND RESULTS

4.1. Graz02 Bike Dataset

To validate our formulation, we perform pixel-wise classification on the Graz02 bikes dataset [23]. We select the first 300 images and use odd numbered images for training and even numbered images for testing. For each image, dense SIFT features are computed from overlapping patches of size 32×32 with a grid spacing of 20 pixels. Beliefs for missing pixels are interpolated from their neighborhoods. Some results are shown in Figure 1.

Table 2 shows that our formulation achieves a better equal-error-rate (EER) on the precision-recall curve than the state-of-the-art in dictionary learning based approaches. Our results match the superpixel based method of [22] which, like

	No CRF	κ_0	κ^*
$\mathbf{D_0}$	55.1	58.2	66.7
$\mathbf{D}^*$	62.3	63.2	72.1

Table 3. EER values on Graz02 bike test set from using **left to right**: no CRF inference, initial CRF weight paramters κ_0, learned CRF weight parameters κ^* and **top to bottom**: initial K-means dictionaries $\mathbf{D_0}$ and dictionaries learned with neighborhood constraints $\mathbf{D}^*$. The benefit of training CRF weight parameters and the use of neighborhood constraints can be seen in isolation. See text for details.

our approach, uses a single scale[2].

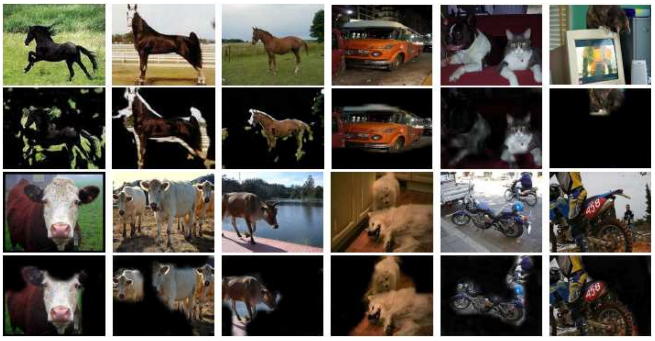

Fig. 2. Some sample results on the Weizmann Horse dataset and VOC 2007 dataset. The advantage of using neighborhood information can be seen for cat segmentation on the cat and dog image in which large patches on both animals are similar and yet inference using our dictionaries was able to extract the cat with rather crisp boundaries.

Benefit of Neighborhood Constraints Table 3 demonstrates in isolation the benefits of training CRF weight parameters and neighborhood constrained learning of dictionaries. Column 1, for instance, shows that dictionaries learned with neighborhood constraints perform better even when inference is carried out without spatial propagation of labels and row 2 generally shows that our learning formulation gives around 6% improvement over the initial dictionaries. Similarly, column 3 shows that learning of CRF weights results in around 10% improvement.

4.2. VOC 2007

Table 4 presents the EER values for figure-ground segmentation on the 20 categories of the Pascal VOC 2007 dataset [24] and compares with the performance of dictionaries learned using KSVD [10]. Training and testing is performed on the images containing the relevant category. Figure 2 shows some sample results. The advantage of using neighborhood information can be seen for cat segmentation on the cat and dog image in which large patches on both animals are similar and yet inference using our dictionaries was able to extract the cat with rather crisp boundaries.

[2]Even single scale superpixels offer more scale information compared to fixed size patches on fixed grids

Class	KSVD [10]	Ours
aeroplane	35.2	**43.7**
bicycle	28.3	**41.2**
bird	35.3	**42.3**
boat	26.3	**35.5**
bottle	16.1	**30.2**
bus	43.7	**69.0**
car	29.1	**43.2**
cat	39.9	**63.3**
chair	9.1	**10.6**
cow	46.0	**70.0**
dining table	38.8	**52.7**
dog	33.3	**51.5**
horse	36.6	**42.0**
motorbike	47.2	**62.9**
person	28.3	**43.0**
potted plant	23.0	**31.4**
sheep	47.5	**54.3**
sofa	21.8	**28.0**
train	54.3	**74.0**
tv/monitor	16.3	**29.1**

Table 4. EER values for figure-ground segmentation on the VOC 2007 dataset.

For classification against all other categories in the manner of Yang & Yang [12], we trained a dictionary for the cow category on the 422 training images and tested on all 210 test images. We obtain 8.5% EER on the pixel level compared to the 8% on patches reported in [12]. It should be noted that in [12], going from patch to pixel level was seen to decrease performance by around 10%.

5. CONCLUSION

We have introduced a novel discriminative dictionary learning procedure that imposes neighborhood contraints during the learning process. This is motivated by the smoothness and boundary-preserving priors on natural images and achieved by embedding dictionary learning in a CRF framework. As an additional benefit, such smoothness constraints lead to stable dictionary learning which is inherent to the problem of discriminative dictionary learning. Detailed analysis on the Graz02 bike dataset demonstrates a distinct quantitative as well as qualitative advantage over competing dictionary-based approaches.

While results are shown for the 2-class case only, the formulation applies to the general N-class case. However, this can potentially lead to a significant increase in sparse coding computation. An alternative is an N-class learning formulation that performs discriminative sparse coding on a single dictionary for all classes.

An interesting extension is the use of sparse long range random fields [25] for dictionary learning via multiscale information.

6. REFERENCES

[1] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman, "Discriminative learned dictionaries for local image analysis.," in *CVPR*, 2008.

[2] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman, "Supervised dictionary learning," in *NIPS*, 2008.

[3] Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features.," in *CVPR*, 2010.

[4] Nazar Khan and Marshall Tappen, "Stable discriminative dictionary learning via discriminative deviation," in *ICPR*, 2012.

[5] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto, "Localizing objects with smart dictionaries," in *ECCV*, 2008.

[6] Zhuolin Jiang, Zhe Lin, and Larry S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *CVPR*, 2011, pp. 1697–1704.

[7] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[8] Jingen Liu and Mubarak Shah, "Learning human actions via information maximization," in *CVPR*, 2008.

[9] Jingen Liu, Yang Yang, and Mubarak Shah, "Learning semantic visual vocabularies using diffusion distance," in *CVPR*, 2009.

[10] Michal Aharon, Michael Elad, and Alfred Bruckstein, "K-svd: Design of dictionaries for sparse representation," in *SPARS*, 2005.

[11] Huimin Guo, Zhuolin Jiang, and Larry S. Davis, "Discriminative dictionary learning with pairwise constraints," in *ACCV*, 2012.

[12] Jamie Yang and Ming-Hsuang Yang, "Top-down visual saliency via joint CRF and dictionary learning.," in *CVPR*, 2012.

[13] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online dictionary learning for sparse coding.," in *ICML*, 2009, vol. 382.

[14] Wei Dai, Tao Xu, and Wenwu Wang, "Simultaneous codeword optimization (simco) for dictionary update and learning," *CoRR*, vol. abs/1109.5302, 2011.

[15] J. Andrew Bagnell and David M. Bradley, "Differentiable sparse coding," in *NIPS*, 2008, pp. 113–120.

[16] Jianchao Yang, Kai Yu, and Thomas S. Huang, "Supervised translation-invariant sparse coding," in *CVPR*, 2010, pp. 3517–3524.

[17] Sanjivv Kumar and Martial Herbert, "Discriminative random fields: A discriminative framework for contextual interaction in classification.," in *ICCV*, 2003.

[18] J Lafferty, McCallum A, and Pereira F, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data.," in *ICML*, 2001.

[19] S. V. N. Vishwanathan, Nicol N. Schraudolph, Mark W. Schmidt, and Kevin P. Murphy, "Accelerated training of conditional random fields with stochastic gradient methods," in *In ICML*, 2006, pp. 969–976.

[20] Pushmeet Kohli, Lubor Ladicky, and Philip H. S. Torr, "Robust higher order potentials for enforcing label consistency," in *CVPR*, 2008.

[21] Marcin Marszalek and Cordelia Schmid, "Accurate object recognition with shape masks," *International Journal of Computer Vision*, pp. 191–209, 2012.

[22] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proc. ICCV*, 2009.

[23] Andreas Opelt, Axel Pinz, Michael Fussenegger, and Peter Auer, "Generic object recognition with boosting," *PAMI*, vol. 28, 2004.

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," .

[25] Yunpeng Li and Daniel P. Huttenlocher, "Sparse long-range random field and its application to image denoising.," in *ECCV (3)*, 2008, pp. 344–357.