Multiclass Discriminative Fields for Parts-Based Object Detection

Sanjiv Kumar and Martial Hebert
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{skumar,hebert}@ri.cmu.edu

Abstract

In this paper, we present a discriminative framework for parts-based object detection based on the multiclass extensions of binary discriminative fields described in [1]. These fields allow simultaneous discriminative modeling of the appearance of individual parts and the geometric relationship between them. The conventional Markov Random Field (MRF) formulations cannot be used for this purpose because they do not allow the use of data while modeling interaction between labels which is crucial for enforcing geometric consistencies between parts. The proposed technique can handle object deformations, occlusions and multiple-instance detection in a single trained model with no added computational efforts. The parameters of the field are learned using efficient maximum marginal approximations and inference is carried out using loopy belief propagation. We demonstrate the efficacy of this approach through controlled preliminary experiments on rigid and deformable synthetic toy objects.

1 Introduction

Object detection has been a long standing problem in computer vision. Even though several promising approaches have been proposed in the literature, generic category-level object detection under complex variations in appearances, object deformations and occlusions is still is a challenging problem. In this paper we propose a discriminative technique based on multiclass extensions of the binary Discriminative Random Fields (DRFs) [1] to address this challenge. The proposed framework has three key advantages: First, during classification, it probabilistically enforces the appearance of individual parts and geometric consistency between parts simultaneously making the classification robust to ambiguities and deformations. Second, the part appearances as well as the relations between parts are modeled using local discriminative models, thus avoiding the need of learning generative models which may be hard to learn for complex data. Last, the final classification is obtained using efficient inference over the graph carried out using existing techniques without needing exhaustive search in the solution space.

Several detection techniques try to detect the object as a 'whole' by classifying a window scanned over the image if it contains the object of interest [2][3][4]. Even though these approaches have been successfully applied to detect faces and cars etc., they tend to

have problems when objects are occluded, or when they undergo significant deformations or articulations. As will be explained later, these issues can be handled naturally in the multiclass DRF framework without needing any extra modeling or computational effort.

The 'parts-based' approach to object detection is based on the idea of identifying 'characteristic' parts of the object in the image. The parts-based techniques that first detect the object parts purely on the basis of their appearance and then refine these part detections using geometric reasoning [4][5] may yield inaccurate results if the appearance of the parts in images is noisy or ambiguous. So, it is desirable to have techniques that detect the parts not only on the basis of their individual appearance but also by enforcing geometric relationship of the parts *simultaneously*. This can be achieved by interpreting the part detection as a labeling problem in which labels (i.e. parts) of the object are dependent on other labels. Thus, this problem can be viewed as a classification problem in a *random field* framework. This idea forms the basis of this paper.

Note that there exist other promising parts-based techniques, which view the detection problem as an explicit search over the image parts [6][7]. A graph is formed over the object parts which allows one to model appearance and relations between the parts simultaneously. But the final classification is carried out by searching the solution space which is $O(N^P)$ problem where N is the number of total parts in the image, and P is the number of object parts. For computational tractability, N and P are restricted to be small (typical choice is 20 for N and 5 for P). On the other hand our DRF based approach defines a graph over the image sites and detection task is seen as labeling individual image parts. At classification time, this has a computational complexity of $O(NP^2)$ which allows efficient inference even if N is in hundreds as we will show later in experimental results.

Also, the graph based techniques usually detect a single instance of the object in the scene. To detect multiple instances of objects in the scene, either the number of instances should be known a-priori or a threshold will be required on the candidate scores. On the other hand, the DRF based framework allows detection of multiple instances naturally without needing any such information. Finally, all the graph based techniques of object detection work exclusively in generative framework which may spend a lot of resources on modeling the generative models for complex part appearances and part relations which are not particularly relevant to final classification task. Moreover, learning realistic class density models may become even harder when the training data is limited. To the best of our knowledge this paper presents the first graphical model based approach to object detection that models the part appearances and their geometric relations in a discriminative framework.

1.1 Approach

In the parts-based paradigm of object detection, given generic parts in the image, our aim is to label each part whether it is a 'specific' part of the object or it is 'background'. We will call each part in the image as an image site. Let the observed data from an input image be given by $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i \in S}$ where $\boldsymbol{y}_i$ is the data from i^{th} site, $\boldsymbol{y}_i \in \Re^c$, and S is the set of all the image sites (i.e. parts). Let the corresponding labels at the image sites are given by $\boldsymbol{x} = \{x_i\}_{i \in S}$, where $x_i \in \{1, \ldots, C\}$ and C is the number of classes. Suppose first $(C-1)$ labels correspond to specific object parts and the C^{th} label corresponds to the background class.

In [1] we presented a binary Discriminative Random Field (DRF) for image analysis, which is a type of Conditional Random Field (CRF) proposed by Lafferty et al. [8], which allows to model arbitrarily complex dependencies in the observed data as well as the labels in a principled manner. These fields discriminatively model the conditional distribution of the labels given the observed data, $P(\boldsymbol{x}|\boldsymbol{y})$, directly as a Markov Random Field (MRF). In this work, we propose to extend the binary DRFs to allow an image site to take multiple class labels as required for parts-based object detection. In addition, the extended DRFs work on

graphs where topology is not fixed to be the same for all images, since in object detection task, sites are not restricted to a regular grid type configuration on a 2D lattice.

2 Multiclass Discriminative Random Field

Assuming only up to pairwise clique potentials to be nonzero in the graph, the distribution over the labels $\boldsymbol{x}$ given the observations $\boldsymbol{y}$ in CRFs can be written as,

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z} \exp\left(\sum_{i \in S} A(x_i, \boldsymbol{y}) + \sum_{i \in S} \sum_{j \in \mathcal{N}_i} I(x_i, x_j, \boldsymbol{y}) \right) \tag{1}$$

where Z is a normalizing constant known as the partition function, $\mathcal{N}_i$ is the set of neighbors of site i. Here $A(x_i, \boldsymbol{y})$ and $I(x_i, x_j, \boldsymbol{y})$ are the unary and pairwise potentials called association and interaction potentials respectively. In the DRF framework, $A(x_i, \boldsymbol{y})$ is viewed as a term that outputs the association of the site i with class x_i which is modeled using a local discriminative model. This view allows one the flexibility to choose domain-specific discriminative classifiers suitable for specific task domains.

Generalizing the binary form of association potential, $A(x_i, \boldsymbol{y})$ in multiclass DRF is modeled as,

$$A(x_i, \boldsymbol{y}) = \sum_{k=1}^{C} \delta(x_i = k) \log P(x_i = k|\boldsymbol{y}) \tag{2}$$

where, $\delta(x_i = k)$ is 1 if $x_i = k$ and 0 otherwise. For each site i, let $\boldsymbol{f}_i(\boldsymbol{y})$ be a function that maps the observations $\boldsymbol{y}$ on a feature vector such that $\boldsymbol{f}_i : \boldsymbol{y} \rightarrow \Re^l$. To extend the local discriminative classifier to induce a nonlinear decision boundary in the feature space, a transformed feature vector at each site i is defined as, $\boldsymbol{h}_i(\boldsymbol{y}) = [1, \phi_1(\boldsymbol{f}_i(\boldsymbol{y})), \ldots, \phi_R(\boldsymbol{f}_i(\boldsymbol{y}))]^T$ where $\phi_r(.)$ are arbitrary nonlinear functions. The first element of the transformed vector is kept as 1 to accommodate the bias parameter. Note that in the case of object detection, the vector $\boldsymbol{h}_i(\boldsymbol{y})$ encodes the appearance based features of the i^{th} site (or part). To model $P(x_i = k|\boldsymbol{y})$, in this paper we will simply use the multiclass version of the logistic form we chose for the binary DRFs in our previous work [1]. This leads to the softmax function in the multiclass case,

$$P(x_i = k|\boldsymbol{y}) = \begin{cases} \frac{\exp(\boldsymbol{w}_k^T \boldsymbol{h}_i(\boldsymbol{y}))}{1 + \sum_{l=1}^{C-1} \exp(\boldsymbol{w}_l^T \boldsymbol{h}_i(\boldsymbol{y}))} & \text{if } k < C \\ \\ \frac{1}{1 + \sum_{l=1}^{C-1} \exp(\boldsymbol{w}_l^T \boldsymbol{h}_i(\boldsymbol{y}))} & \text{if } k = C \end{cases} \tag{3}$$

Here $\boldsymbol{w}_k$ are the model parameters for $k = 1 \ldots C-1$. For a C class classification problem, one needs only $C-1$ independent hyperplanes which may lie in a high dimensional (kernel-projected) space inducing a non-linear decision boundary in the original feature space. Note that this choice of $P(x_i = k|\boldsymbol{y})$ leads to the unary potential which is linear in parameters similar to the CRFs given in [8] with a subtle difference that the parameters $\boldsymbol{w}_k$, for $k = C$, are set to $\mathbf{0}$. Note that other domain-specific choices of $P(x_i = k|\boldsymbol{y})$ are also possible. In the application of object detection, the association potential models the individual appearance of each part in the image.

The interaction potential in DRFs predicts how the labels at two sites interact given the observations. Generalizing the interaction potential given for binary DRFs in [1], interaction potential for multiclass DRFs can be written as,

$$I(x_i, x_j, \boldsymbol{y}) = \sum_{k=1}^{C} \sum_{l=1}^{C} \boldsymbol{v}_{kl}^T \boldsymbol{\mu}_{ij}(\boldsymbol{y}) \delta(x_i = k) \delta(x_j = l) \tag{4}$$

Here, $\boldsymbol{\mu}_{ij}(\boldsymbol{y})$ is the pairwise relational vector for a site pair (i,j), and $\boldsymbol{v}_{kl}$ are the model parameters. Note that in the case of object detection, vector $\mu_{ij}(\boldsymbol{y})$ encodes the pairwise features required for forcing geometric and possibly photometric consistency in the pair of parts. For undirected graphs, the site pairs are unordered sets implying that $\boldsymbol{v}_{kl} = \boldsymbol{v}_{lk}$ for $k,l = 1 \ldots C$. This form of interaction potential given in (4) is a generalization of commonly used Potts model, which can be recovered from (4) if $\boldsymbol{v}_{kl} = \mathbf{0}$ when $k \neq l$, and all the elements of the vector $\boldsymbol{v}_{kl}$ are set to zero except the bias term when $k = l$. Similar to the interaction potential of the binary DRF, multiclass interaction potential can be seen as a pairwise discriminative model which partitions the pairwise relational feature space (induced by the features $\mu_{ij}(\boldsymbol{y})$) in $C(C+1)/2$ regions.

It is important to note that, to enforce the geometric consistency relationship between parts, the interaction between part labels has to use observed data (e.g. the location of patches). Since, in DRFs, the pairwise potential I is a function of observed data, these fields allow a principled way of solving the detection problem in a random-field framework. On the contrary, in the conventional MRFs, the conditional distribution over labels is modeled as $P(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})$, where $P(\boldsymbol{x})$ is used for modeling the label interaction. Since $P(\boldsymbol{x})$ does not allow use of data $\boldsymbol{y}$ while modeling label interactions, conventional forms of MRFs cannot model the geometric consistency simultaneously with appearance.

3 Parameter learning and inference

Let θ be the set of DRF parameters where $\theta = \{\{\boldsymbol{w}_k\}_{k=1\ldots C-1}, \{\boldsymbol{v}_{kl}\}_{k,l=1\ldots C}\}$. Given M i.i.d. labeled training images, the maximum likelihood estimates of the parameters are given by maximizing the log-likelihood $l(\theta) = \sum_{m=1}^{M} \log P(\boldsymbol{x}^m|\boldsymbol{y}^m,\theta)$. To learn the parameters using gradient ascent, the derivative of the log-likelihood, after some algebraic manipulations, can be written as,

$$\frac{\partial l(\theta)}{\partial \boldsymbol{w}_k} = \sum_m \sum_{i \in S^m} \Big(\delta(x_i^m = k) - \langle \delta(x_i = k) \rangle\Big) \boldsymbol{h}_i(\boldsymbol{y}^m) \tag{5}$$

$$\frac{\partial l(\theta)}{\partial \boldsymbol{v}_{kl}} = \sum_m \sum_{i \in S^m} \sum_{j \in \mathcal{N}_i} \Big(\delta(x_i^m = k)\delta(x_j^m = l) - \langle \delta(x_i = k)\delta(x_j = l) \rangle\Big) \boldsymbol{\mu}_{ij}(\boldsymbol{y}^m) \tag{6}$$

Here $\langle . \rangle$ denotes expectation with respect to the distribution $P(\boldsymbol{x}|\boldsymbol{y}^m,\theta)$. Generally the expectation in (5) and (6) cannot be computed analytically even for moderately sized problems due to the combinatorial number of elements in the configuration space of labels $\boldsymbol{x}$. One possible way of approximating the expectations is to use sampling procedures, e.g. Markov Chain Monte Carlo (MCMC). But the main problem with MCMC techniques is that to sample from the stationary distribution one needs to wait for 'burn-in' period which is to determined empirically. In addition to being slow, MCMC techniques usually yield estimates with high variance [9]. Alternatively one could estimate expectations using pseudo-marginals returned by loopy Belief Propagation (BP) [10] or any other technique e.g. Expectation Propagation (EP) etc.

In this work we use BP to get the pseudo-marginals. However, we used a slightly different approximation of the gradient which uses maximum marginals label estimates. Let, $\tilde{x}_i = \arg\max_{x_i} P_i(x_i|\boldsymbol{y}^m,\theta)$. Then the expectations in (5) and (6) are approximated as:

$$\langle \delta(x_i = k) \rangle = \delta(\tilde{x}_i = k) \quad \text{and} \quad \langle \delta(x_i = k)\delta(x_j = l) \rangle = \delta(\tilde{x}_i = k)\delta(\tilde{x}_j = l)$$

This approximation is equivalent to approximating the partition function Z by concentrating all the its mass on a single point given by thresholded stationary points of the Bethe free energy. With this approximation, the resulting gradient descent updates resemble the

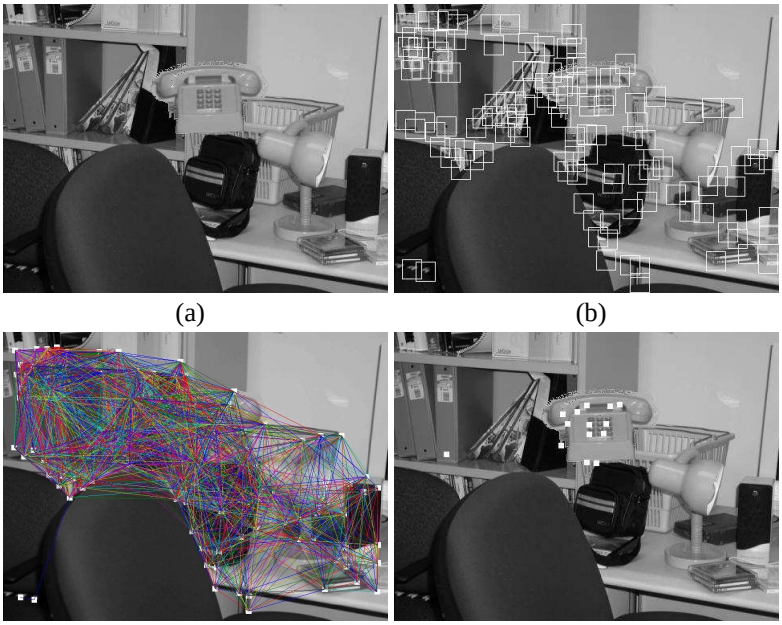


Figure 1: Detection of a rigid object (phone) embedded in a cluttered scene. (a) Input image. (b) Extracted parts. (c) Graph joining parts with their neighbors. (d) Detection results. Parts that are classified as object parts are shown highlighted.

perceptron learning rules with interesting weak convergence properties which are being explored in detail in our other work [11]. Empirically we found that this approximation gave comparative performance to the pseudo-marginal approximation but was found to be much faster.

For inference, in this work, we used sum-product version of loopy BP to find the Maximum Posterior Marginal (MPM) estimates of the labels on the image sites, which are optimal for sitewise zero-one loss function. In fact, as being studied in a separate work [11], the duality of using the same approximation for parameter learning and inference tends to minimize the classification error .

4 Object detection experiments

To test the performance of multiclass DRFs, we conducted preliminary experiments with object detection on synthetic data. The aim of these controlled experiments was to illustrate the detection framework under discriminative fields and verify their performance under ambiguities, deformations and occlusions.

Let us denote all the pairs of parts that have one or both the parts from the background by 'background pairs' and the rest by 'object pairs'. The label C denotes the background class. To separate the background pairs from the object pairs using a single hyperplane we must have $\boldsymbol{v}_{kl} = \boldsymbol{v}_b$ if $k = C$ or $l = C$ or both. Further, without the loss of generality $\boldsymbol{v}_b$ can be set to $\mathbf{0}$, since to partition a K class problem we need to learn only $K-1$ independent hyperplanes.

4.1 Experiments with rigid object

In the first set of experiments the aim was to (a) illustrate the detection framework under DRFs using a rigid object, (b) verify the performance under object occlusions, and (c)

validate the capability of the framework to deal with multiple objects in the scene. In these experiments, the task was to detect a phone in a cluttered scene (Fig. 1(a)). Synthetic training and test data was generated by taking a mask of the phone and embedding its affine distortions in 300 random office backgrounds. $\pm 10\%$ percent variation was allowed in scale and shear. For each training image, at first, interest points were detected using the Harris corner detector and a patch of size 25×25 pixels around each interest point was called a part as shown in Fig. 1(b). A graph was generated using these patches as nodes as shown in Fig. 1(c). All patches within a specified radius from a patch were called neighbors of that patch. Note that the resulting graph is no longer a regular grid lattice and that each node in the graph will usually have different number of neighbors. In this work, we used a uniform distribution over the graph structures which leads to 'averaging' over all the graphs in the training images. We intend to explore in the future if better distributions could be learned over the graph structure itself.

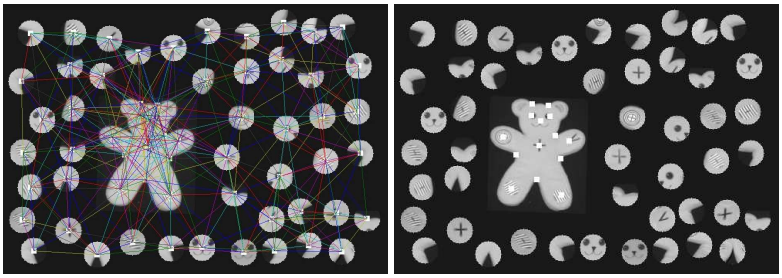
The appearance based features used in the association potential, $\boldsymbol{f}_i(\boldsymbol{y})$, were computed based on the gradient orientation histograms weighted by the gradient magnitude and quadratic transformations were used to get $\boldsymbol{h}_i(\boldsymbol{y})$. The pairwise features, $\boldsymbol{\mu}_{ij}(\boldsymbol{y})$, were just the distances between the parts. In the future, joint appearance may also be added. For this problem, the number of classes, C, was fixed to 17 based on the object part-detector output while training. The model had overall 3230 parameters which were learned successfully using the BP-based learning technique described in Section 3. The associations parameters $\boldsymbol{w}_k$ were initialized from the softmax classifier parameters, while the interaction parameters $\boldsymbol{v}_{kl}$ were initialized at 0. At the test time, BP was used to infer the optimal labeling of the parts. In Fig. 1(d) all the parts that were labeled as any of the object parts are shown highlighted. To generate the final object hypothesis, one may use simple postprocessing step (e.g. location based clustering) to filter any isolated false positives. Training took about 50 iterations and two hours, while the average time taken for inference was 1.35 sec per image on a 2GHz machine.

To demonstrate the effect of occlusion, we synthetically blocked the right half of the phone and the DRF detection results are shown in the left image in Fig. 2. To verify multiple instance detection under this framework, two affine distorted versions of the phone were embedded randomly in the scene and the corresponding detection results are shown in the right image in Fig. 2. Note that no information about number of objects in the scene was known, and the same learned model described in the previous paragraph was used for detection in both experiments.

Figure 2: Toy examples constructed to demonstrate detection with occlusion (left), and with multiple object instances in the scene (right) using the same learned model.

4.2 Experiments with deformable object

In the second set of synthetic experiments, we explored the answers to two questions: First, can the DRF model learn all the deformations of a deformable object in a single model, and



(a) (b)

Figure 3: Detection of a deformable object (teddy) in a synthetic scene in which the object patches are inserted as background patches to confuse the appearance based detection. (a) Graph over input image joining patches with their neighbors. (b) Detection results. Patches that are classified as object parts are shown highlighted. Note that DRF was able to ignore the background patches.

second, can it automatically learn to trade off the appearance with the geometric constraints between parts in the presence of ambiguities?

For this, we constructed an experimental setup with an articulated toy object shown in (Fig. 3) in which different joints of the object could be deformed independently. The training and test sets were generated by embedding affine distortions of different deformed versions of the object in synthetic backgrounds. To confuse the appearance, we randomly inserted the object patches in the background (Fig. 3). Clearly, if appearance alone were used to classify the parts, everything would be classified as background. This is because there are many more background patches than the object patches in the training set and a discriminative classifier will try to reduce the classification error by simply assigning all the object patches to the background class. However, the geometric relationship along with the appearance should be able to restrict the choice of parts being from the object. This is exactly what is exploited by the DRF as shown by the result in Fig. 3(b).

Some more results on different deformations of the object are given in Fig. 4. To compare the DRF results with just appearance based detector (softmax classifier), as expected, softmax assigned all the 4287 object parts incorrectly as the background, while the DRF was able to classify with high accuracy 4258 object parts as the object. The background detection was the same for both of them. Note that for all the affine and articulated deformations in the object, only a single DRF was learned to account for all these variations. The training needed about 50 iterations and less than one hour while the testing took on an average 0.24 sec to process each image on a 2 GHz machine.

5 Conclusion and future work

In this work, we have presented a new discriminative paradigm for deformable object detection which can simultaneously model individual part appearances and their geometric consistency. This is possible in the DRF framework in a random-field setting since the discriminative fields allow the use of observed data in pairwise potentials. The proposed framework can handle deformations, occlusions and multiple-instance detection using a single learned model without needing any extra computational efforts. Also, we have shown that it is possible to do efficient parameter learning and inference over such models, without needing exhaustive search. The preliminary experiments were conducted as a proof-of-concept to demonstrate DRF advantages. Clearly, the next important step is to apply this framework to the real-world detection tasks and compare its performance with existing techniques. Scale invariance can be achieved in this framework by choosing scale invariant unary and pairwise features, or by using the cliques of size three or more in the

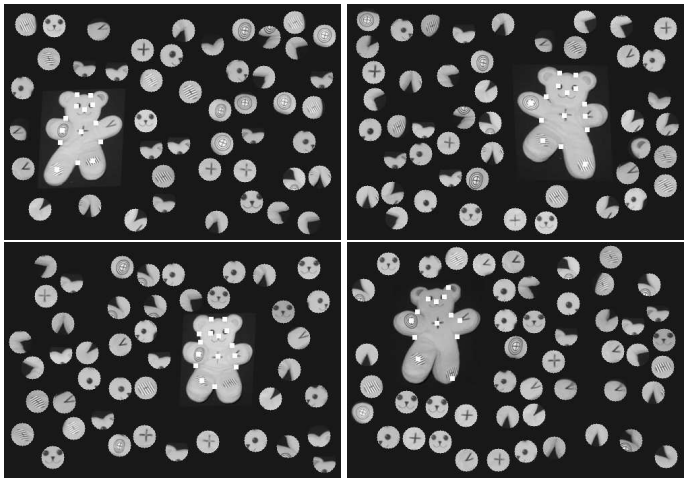


Figure 4: Synthetic detection experiments with various object deformations. Note that for all the affine and articulated deformations in the object, only a single DRF was learned to account for all these variations.

model.

Acknowledgments

Our thanks to J. Lafferty and J. August for helpful discussions on conditional fields, and T. Naik and A. Gallagher for data collection.

References

[1] S. Kumar and M. Hebert. Discriminative fields for modeling spatial dependencies in natural images. *in advances in Neural Information Processing Systems (NIPS)*, December 2003.

[2] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 00)*, 2000.

[3] Paul Viola and Michael Jones. Robust real-time object detection. *in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 01)*, 2001.

[4] Shyjan Mahamud and Martial Hebert. Minimum risk distance measure for object recognition. *in Proc IEEE International Conference on Computer Vision (ICCV 03)*, 2003.

[5] David G. Lowe. Object recognition from local scale-invariant features. *in Proc. International Conference on Computer Vision (ICCV 99)*, pages 1150–1157, 1999.

[6] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. *In Proc.IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 00)*, June 2000.

[7] Rob Fergus, Pietro Perona, and Andrew Zisserman. Object class recognition by unsupervised scale-invariant learning. *In Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 03)*, 2:264–271, 2003.

[8] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In Proc. Int. Conf. on Machine Learning*, 2001.

[9] G. E. Hinton. Training product of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

[10] B. Frey and D. J. C. Mackay. A revolution: Belief proapagation in graphs with cycles. *In Proc. Advances in Neural Information Processing Systems*, 10, 1997.

[11] S. Kumar and M. Hebert. Approximate parameter learning in discriminative fields. *Snowbird Learning Workshop*, 2004.