Discriminative maximum margin image object categorization with exact inference

Qinfeng Shi
Australian National University

Luping Zhou
University of Wollongong, lupingz@uow.edu.au

Li Cheng
Toyota Technological Institute

Dale Schuurmans
University Of Alberta

Follow this and additional works at: https://ro.uow.edu.au/eispapers

Part of the Engineering Commons, and the Science and Technology Studies Commons

Recommended Citation
Shi, Qinfeng; Zhou, Luping; Cheng, Li; and Schuurmans, Dale, "Discriminative maximum margin image object categorization with exact inference" (2010). Faculty of Engineering and Information Sciences - Papers: Part A. 1855.

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au
Discriminative maximum margin image object categorization with exact inference

Abstract
Categorizing multiple objects in images is essentially a structured prediction problem: the label of an object is in general dependent on the labels of other objects in the image. We explicitly model object dependencies in a sparse graphical topology induced by the adjacency of objects in the image, which benefits inference, and then use maximum margin principle to learn the model discriminatively. Moreover, we propose a novel exact inference method, which is used in training to find the most violated constraint required by cutting plane method. A slightly modified inference method is used in testing when the target labels are unseen. Experiment results on both synthetic and real datasets demonstrate the improvement of the proposed approach over the state-of-the-art methods.

Keywords
margin, maximum, discriminative, exact, inference, categorization, object, image

Disciplines
Engineering | Science and Technology Studies

Publication Details

This conference paper is available at Research Online: https://ro.uow.edu.au/eispapers/1855
Discriminative Maximum Margin Image Object Categorization with Exact Inference

Qinfeng Shi  
ANU and NICTA  
Australia  
qinfeng.shi@anu.edu.au

Luping Zhou  
ANU and NICTA  
Australia

Li Cheng  
TTI-Chicago  
USA

Dale Schuurmans  
Univ of Alberta  
Canada

Abstract

Categorizing multiple objects in images is essentially a structured prediction problem: the label of an object is in general dependent on the labels of other objects in the image. We explicitly model object dependencies in a sparse graphical topology induced by the adjacency of objects in the image, which benefits inference, and then use maximum margin principle to learn the model discriminatively. Moreover, we propose a novel exact inference method, which is used in training to find the most violated constraint required by cutting plane method. A slightly modified inference method is used in testing when the target labels are unseen. Experiment results on both synthetic and real datasets demonstrate the improvement of the proposed approach over the state-of-the-art methods.

1. Introduction

A fundamental problem of image understanding is to categorize objects in images based on visual content. This is inherently a structured classification problem for object segments, which is usually obtained by image segmentation or human interaction. Recently, discriminative learning methods have been employed for this task, e.g. using conditional random fields (CRFs) [6–8], to exploit the spatial Markovian dependencies between neighboring pixels. A key requirement of these approaches however is an inference procedure for finding the best label assignment given a image content and the model. Previous approaches either ignore dependencies between the objects within an image, (which boils down to independent identically distributed (i.i.d) classification such as SVMs), or perform approximate inference either by message passing methods e.g. loopy belief propagation (LBP) [13, 14] or by graph-cuts [1, 5]. However, both graph-cut and LBP are approximate inference methods for multi-class problems in loopy graphs.

In this paper, we propose a discriminative approach for image object recognition based on the maximum margin principle [12]. This approach allows us to explicitly model the object-based nature of the problem by incorporating features and relations of segments rather than pixels. In particular, as an alternative of junction tree algorithm, we develop a novel exact inference algorithm for obtaining the global optimal assignment in our set-up, even in the multi-class scenario.

Related Work Recently, Corso et al. [2] have proposed the graph-shift algorithm and applied the CRF learning framework for image labeling, aiming to dynamically update the parent-child relationship in a hierarchical decomposition of a image. This is however still an approximate inference method. Zhu et al. [15] introduce an and/or graph, a special context sensitive grammar, and propose to infer both the label and the latent grammar graph from an image in an unsupervised manner. Meanwhile, Petrov and Klein [9] propose to discriminatively learn the log-linear hierarchical models and the grammars. As a trade-off, their objective function becomes non-convex due to the introduction of the latent variables, therefore there is no guarantee for global optimality.

However models which take into account dependence of segments have only recently been introduced. [10] modeled the problem with a fully connected CRF where nodes are the objects and the edges encode the co-occurrence counts of labels in the training set. Learning is performed by approximate Maximum-Likelihood estimation through sampling, since exact inference is intractable. Performance improvement is reported when compared against independent predictions per object. This model was extended in [4] by incorporating relative location information, e.g. sky it is typically above ground.

2. Modeling Joint Categorization

Problem Formulation We assume that a segmentation of the original image is given. This can be done by sev-
eral means, such as shown in [4]. We assume each segment has a latent label from a fixed dictionary of \( L \) labels \( \mathcal{L} = \{l_1, \ldots, l_L\} \). The goal is to infer the category of each segment. We model an arbitrary segmentation (a partition) of an image in \( N \) parts as a random vector \( X = (X^1, \ldots, X^N) \). A segmented image is therefore a specific realization of such random vector, which we denote by \( x = (x^1, \ldots, x^N) \). Since we assume a segmentation as given, this random vector is always observed. In practice each segment \( x^i \) will have a feature vector associated to it, which will be incorporated into our model. We denote an arbitrary joint labeling of the segmented image \( x \) by a random vector \( Y = (Y^1, \ldots, Y^N) \). A realization of \( Y \) is a specific joint labeling, which we denote by \( y = (y^1, \ldots, y^N) \), where \( y^i \in \mathcal{L}, \forall i, 1 \leq i \leq N \).

The Model We cast this estimation problem as finding a discriminant function \( f(x, y) \) such that for an image \( x_t \) with objects \( (x^1_t, \ldots, x^N_t) \), we assign the categories that receive the best score with respect to \( f \).

\[
y^* = \arg\max_y f(x_t, y).
\] (1)

As in many learning algorithms, we consider functions that are linear in some feature representation

\[
f(x, y; w) = \langle w, \phi(x, y) \rangle.
\] (2)

Here \( \phi(x, y) \) is a feature map and \( w \) is the corresponding parameter vector. As in CRF, the feature map \( \phi(x, y) \) can be decomposed into nodes and edges:

\[
\phi(x, y) = \sum_i \phi_1(x, y^i) + \sum_{(i,j) \in A} \phi_2(x, y^i, y^j).
\] (3)

Here \( \phi_1 \) is the node feature or the intra-object feature and \( \phi_2 \) is the edge feature or inter-object feature. More details of features are described in section 3.

2.1. Maximum Margin Training

We now present a maximum margin training for categorizing image object. The set of labeled segments in one image is an instance. One of the advantages of this method is its ability to incorporate the cost function that the classifier is evaluated with. Let \( \Delta(y_t, y^*) \) to be the cost of predicting \( y^* \) instead of the true label \( y_t \). Typically one can choose \( \Delta(y_t, y^*) \) to be the hamming loss. And we follow the general framework of [12] and look for model parameter \( \omega \) that separates the true label \( y_t \) from the other \( y \) with some margin depending on \( \Delta \) additively.

\[
\min_{w, \xi} \lambda \frac{\|w\|^2}{2} + \frac{1}{T} \sum_{t=1}^T \xi_t
\]

s.t. \[ \langle w, \phi(x_t, y_t) \rangle - \langle w, \phi(x_t, y) \rangle \geq \Delta(y_t, y) - \xi_t, \]

\[ \xi_t \geq 0, \forall t = 1, \ldots, T, \forall y \in \mathcal{Y}, \]

(4)

where \( T \) is the number of training images and \( \lambda \) is the regularization constant which is usually determined by model selection.

This problem (4) is intractable, since the configuration space \( \mathcal{Y} \) is exponentially large. However, [12] shows that this problem can be approximately solved in polynomial time with good precision. The key idea is to find the most violated constraints for the current set of parameters and satisfy them up to some precision. In order to do this, one needs to find

\[
y^* = \arg\max_{y \in \mathcal{Y}} \Delta(y_t, y) + f(x_t, y; w),
\] (5)

using our inference method. Similarly, in the prediction phase when the true label \( y_t \) is not accessible, we would infer \( y^* \) by

\[
y^* = \arg\max_{y \in \mathcal{Y}} f(x, y; w).
\] (6)
Figure 2: An exemplar illustration of performing back-tracking on the graph of Figure 1. The boxes with solid boundary are the subgraphs, of which the classes are assigned. The top layer box represented by blue $G_+$ decomposes to blue $G_-$ and $G_+ \setminus \cdot$. And then the blue $G_-$ is considered as a new $G_+$ in red color and it starts decomposing again to red $G_-$ and $G_+ \setminus \cdot$.

2.2. Exact Inference

We now propose a novel exact inference algorithm for solving (5).

Let $G_\infty$ be the full induced graph of an image $x_t$, $G$ be the set of all possible subgraphs of $G_\infty$. The goal is to find the associated $y$ on nodes that give the highest $S(x_t, y) = \Delta(y_t, y) + f(x_t, y; w)$ for any $x_t$. Clearly enumerating all possible $y$ for all nodes is intractable, because there are exponential many $y$. Alternatively, we can decompose $S(x_t, y)$ to subgraphs. We find the best labels for small subgraphs and then use them to find the best labels for larger subgraphs iteratively until we are done for the full graph. By this procedure, many non-optimal configuration will be discarded. Before going to more detail, we shall introduce some definition and notions. The boundary nodes of a graph $G$ are the nodes adjacent to any node that is not in $G$. For any $G$, we associate a special label called key label to $G$. The key label has to be on one of the boundary nodes of $G$. When there is only one boundary node, the key label uniquely determines the label of the boundary node.

Given a subgraph $G_+$ and assume its key label is $c_+$, the recursion form is:

$$S(G_+, c_+) = \max_{G_- \subset G_+, c_- \in \mathcal{L}} \{S(G_-, c_-) + g(G_+ \setminus G_-, c_+, G_-, c_-)\},$$

where $G_-$ is a subgraph of $G_+$, $c_-$ is the boundary label of $G_-$, and $c_\pm$ is the label of $G_+ \setminus G_-$. Here we restrict that all nodes in $G_+ \setminus G_-$ have to share the same label $c_\pm$. We further restrict $|G_+| = |G_-| + 1$ to reduce the computational load while retaining the optimal assignment unchanged.

Define $g(G_+ \setminus G_-, c_\pm, G_-, c_-)$ as the incremental function to associate the subgraph $G_+ \setminus G_-$ and its label $c_\pm$ with the subgraph $G_-$ and its key label $c_-$. This function will be used to measure the score increment given $(c_\pm, G_-, c_-)$. Given these definitions, we can express $g(.)$ as

$$g(G_+ \setminus G_-, c_\pm, G_-, c_-) = f(x_t, c_\pm) + \Delta(y_t(G_+ \setminus G_-), c_\pm)$$

$$= \langle w_2, \phi_2(x_t, c_\pm, c_\pm) \rangle + \langle w_1, \phi_1(x_t, c_\pm) \rangle + \sum_{v \in G_+ \setminus G_-} \Delta(y_t(v), c_\pm),$$

where $y(G)$ denotes the label of nodes in $G$, and $y(v)$ denotes the label of node $v$.

Algorithm 1 provides the detailed pseudo-code of our inference algorithm, which is able to find the same optimal assignment in (5) one would obtain by naively enumerating possible $y$. (6) can then be solved as a special case by removing $\Delta$. Basically it has two steps: forward computing and back tracking. The back tracking step is used to retrieve the best subgraphs $G_-$, $G_+ \setminus G_-$ and labels $c_-, c_\pm$. Figure 2 illustrates such step for the 6 objects image in Figure 1.
We begin with $G_+ = \{n_1, n_2, n_3, n_4\}$ and look for the best $G_-, c_-, c_+$ in tables B1, B2, B3. $c_\pm$ determines the $y(n_1)$ to be sky in this case. Then consider $G_+ = \{n_2, n_3, n_4\}$ and look up the tables, we get $y(n_3)$ is grass. Keeping doing so we obtain $y$ over all nodes. As we can see in algorithm 1, most computation is in the forward computing part. The complexity of our inference is exponential to the tree-width (maximal clique size). This is the same as junction tree algorithm. However, introducing some heuristic can considerably reduce the complexity to polynomial with a price that the inference may not be exact. Our attempts on speeding up by heuristic lead to poorer experiment result. Our conjecture is cutting plane method relies on exact or very good approximated inference. Even occasional poor approximation will cause early stop of the optimization as we observed in our experiment.

3. Experiments

Datasets We conducted experiments on three datasets: a synthetic dataset and two well-known datasets: the MSRC object categorization dataset\(^1\) and the Corel dataset\(^2\).

Comparison Methods We compared our approach (called SVM-DP) with three state-of-the-art methods:

SVMs multi-class SVMs [3]. Here we use the LIBSVM package \(^3\) for both linear kernel (SVM-L) and RBF kernel (SVM-RBF);

MRF markov random field (MRF) that utilizes graph-cuts [1, 5] for inference;

CRF conditional random field (CRF) for categorization in [4]. Our implementation of [4] uses exact inference instead of sampling to compute the derivative of log-partition function. This is because it is much easier to implement exact inference than to perfectly reproduce the sampling scheme of [4]. Aiming at a fair comparison, all structured algorithms use the same node and edge features with an exception that the implementation of [4] adds context matrices into the existing edge potential.

Features We assume $\varphi_1$ is composed by a tensor product of instance and label feature functions, given by $\varphi_1(x^i, y^i) = \varphi_1(x^i) \otimes y^i$, where $\varphi_1(x^i)$ is the raw node feature depending only on the observed image segment $x^i$. Similarly $\varphi_2(x, y, y^i) = \varphi_2(x^i, x^j) \otimes y^j$, where $\varphi_2(x^i, x^j)$ is the raw edge feature depending only on the observation as well, and $y^j := [y^j y^j]$. $\varphi_1$ and $\varphi_2$ are assembled from

\(^1\) http://research.microsoft.com/en-us/projects/objectclassrecognition/;
\(^2\) http://www.cs.toronto.edu/~hexm/data/corel_subset.mat;
\(^3\) http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Figure 3: Sample images from the synthetic dataset.

$\varphi_1$ We extract a texton feature vector [11] from each patch, hence every pixel is represented by a texton vector. The node feature for an object is the empirical mean of the texton vector of pixels. The raw node feature $\varphi_1(x^i) = [1 \varphi_1(x^i)]$.

$\varphi_2$ We use the mean of the boosted texton probability density [11] of all interior and boundary pixels of the objects as their edge feature. The raw edge feature $\varphi_2(x^i, x^j) = [1 \varphi_2(x^i) \varphi_2(x^j)]$.

3.1. Synthetic Dataset

To show that our approach is capable of dealing with contextual information in images, we build a synthetic image dataset containing 100 images and 5 object categories: sky, land, water, horse, fish. A gallery of sample images are displayed in Figure 3.

In this dataset, the two object categories, namely horses and fish appear in random positions while satisfying the following contextual constraints: fish always stay in the water, and horses stays on land most of time but occasionally touch the water. There are also the three background categories; namely sky, land and water. Each has a different but fixed intensity value, hence these categories are easy separated by almost any reasonable classifier. The challenge comes from the horse and the fish categories, as they are random samples drawn from Gaussian distributions with the same variance (0.01) but slightly different mean values (100 for horse and 101 for fish). The gray level range is [0, 255]).

We use 20-fold cross-validation, and for simplicity of demonstration, we use a simple node feature, namely the empirical mean of pixel intensity of each object. The edge feature is the absolute difference of two adjacent objects’ pixel intensities tensored with the classes. In this dataset, we notice that even using very small portion of data for training, all the comparison algorithms can achieve reasonable performance, hence we use only 5% for training and 95% for testing. Table 1 compares the accuracies of the various methods. Our proposed approach is able to achieve the highest average accuracy, 97.06%, on test data.
shows our tree displays the test accuracy on both the MRSC and Corel datasets. However, the fish and horse categories show that by increasing the training size, the error on training size — we only use 5 images to train. Our approach (SVM-DP) clearly outperforms the others. As expected, the structured learning algorithms again outperform the i.i.d methods (SVMs), by leveraging the edge features $\phi_2$ of the induced graph and making decision jointly. Polar bear is sometimes confused with water (see Table 5) due to their texton features appear very similar in feature space.

\section*{3.2. Real Datasets}

We conducted separate sets of experiments on two real world image datasets. The first is the MSRC Object categorization dataset, where we specifically used the scenery 1 portion that contains 30 images and 7 classes: building, grass, tree, cow, horse, sheep and sky. The second is the Corel dataset, which has 100 images and 7 classes: hippo, polar bear, water, snow, vegetation, ground, and sky.

<table>
<thead>
<tr>
<th>SVM-L</th>
<th>SVM-RBF</th>
<th>MRF</th>
<th>CRF</th>
<th>SVM-DP</th>
</tr>
</thead>
<tbody>
<tr>
<td>78.12</td>
<td>76.81</td>
<td>78.75</td>
<td>&gt; 10 days</td>
<td>97.06</td>
</tr>
</tbody>
</table>

Table 1: A comparison of accuracies on the synthetic dataset using 20-folds cross-validation. The exact inference in fully connected graph in CRF is very expensive when the tree-width is big. As result, CRF can’t finish running after 10 days.

<table>
<thead>
<tr>
<th>T vs. P</th>
<th>sky</th>
<th>land</th>
<th>water</th>
<th>horse</th>
<th>fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>sky</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>land</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>water</td>
<td>0.05</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>horse</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.853</td>
<td>0.145</td>
</tr>
<tr>
<td>fish</td>
<td>0.00</td>
<td>0.00</td>
<td>0.36</td>
<td>0.64</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 2: Confusion matrix (True vs. Predict) of SVM-L on synthetic dataset.

<table>
<thead>
<tr>
<th>T vs. P</th>
<th>sky</th>
<th>land</th>
<th>water</th>
<th>horse</th>
<th>fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>sky</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>land</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>water</td>
<td>0.05</td>
<td>0.00</td>
<td>0.91</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>horse</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>fish</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
<td>0.01</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrix on synthetic dataset using SVM-DP (our approach).

The sky, land and water are relatively easy to classify because their intensity values are distinct. In fact, SVMs do not make any mistakes in predicting these three categories, as shown in Table 2. However, the fish and horse categories turn out to be very difficult for SVMs: Table 2 shows that 45% of horses are misclassified as fish and 36% of fish are misclassified as horses. On the other hand, the structured learning algorithms, including MRF, CRF and the proposed approach, are all capable of capturing the characteristics of the objects and the relationships between pairs of objects, which helps achieve reasonable predictions, even when the node features are less informative. Therefore, they outperform SVMs in this task. For example, Table 3 shows our approach does very well in horse and fish. Note that a small error rate occurs in water category. This is due to small training size — we only use 5 images to train. Our experiment shows that by increasing the training size, the error on water will vanish.

1.5% of the whole image, since these are too small to be meaningful for recognition.

Table 4 displays the test accuracy on both the MRSC and Corel datasets. Our approach (SVM-DP) clearly outperforms the others. As expected, the structured learning algorithms again outperform the i.i.d methods (SVMs), by leveraging the edge features $\phi_2$ of the induced graph and making decision jointly. Polar bear is sometimes confused with water (see Table 5) due to their texton features appear very similar in feature space.

\section*{4. Outlook and Discussion}

In this paper, we propose a principled discriminative method for image object recognition, developed around the large margin principle. In particular, we developed a novel exact inference algorithm that can obtain the optimal assignment of (5) and (6), even in the multi-class scenario. Experiments on synthetic and real-world datasets demonstrate the excellency of the proposed approach.

For future work, a natural extension is to incorporate pixel-wise image segmentation into the proposed framework. Also designing good heuristic procedures to speed up the inference in large tree-width without compromising the performance is worth investigating.

\section*{References}


Figure 4: Examples of images from real datasets. **First Column:** The raw images. **Second Column:** The Categorization result of CRF. **Third Column:** The Categorization result of SVM-DP. **Fourth Column:** The ground truth images. Top two images are from MSRC, and the bottom two images are from Corel. Note that CRF misclassifies the sky as grass on the second image in the second row. The figure is best viewed in color.