Taxonomic Multi-class Prediction and Person Layout using Efficient Structured Ranking

Arpit Mittal, Matthew Blaschko, Andrew Zisserman, Philip H.S. Torr

Taxonomic Multi-class Prediction and Person Layout using Efficient Structured Ranking

Arpit Mittal[1], Matthew B. Blaschko[2], Andrew Zisserman[1], and Philip H. S. Torr[3]

[1] Department of Engineering Science, University of Oxford, UK
[2] Center for Visual Computing, École Centrale Paris, France **
[3] Department of Computing, Oxford Brookes University, UK

Abstract. In computer vision efficient multi-class classification is becoming a key problem as the field develops and the number of object classes to be identified increases. Often objects might have some sort of structure such as a taxonomy in which the mis-classification score for object classes close by, using tree distance within the taxonomy, should be less than for those far apart. This is an example of multi-class classification in which the loss function has a special structure. Another example in vision is for the ubiquitous pictorial structure or parts based model. In this case we would like the mis-classification score to be proportional to the number of parts misclassified.

It transpires both of these are examples of structured output ranking problems. However, so far no efficient large scale algorithm for this problem has been demonstrated. In this work we propose an algorithm for structured output ranking that can be trained in a time linear in the number of samples under a mild assumption common to many computer vision problems: that the loss function can be discretized into a small number of values.

We show the feasibility of structured ranking on these two core computer vision problems and demonstrate a consistent and substantial improvement over competing techniques. Aside from this, we also achieve state-of-the art results for the PASCAL VOC human layout problem.

1 Introduction

Multi-class classification has become of increasing interest as the computational power of computers increases. Standard approaches to multi-class image labeling typically penalize incorrect predictions equally: a motorbike mis-classified as a bicycle receives the same penalty as a motorbike mis-classified as a cow. If taxonomic knowledge is included, a system can be designed to penalize mis-classifications proportional to their taxonomic losses (e.g. the distance along a taxonomic tree, so that for instance misclassifying a car as a van might be

** M. B. Blaschko is also associated with Équipe Galen, INRIA Saclay, Île-de-France, France and Université Paris-Est, LIGM (UMR CNRS), Center for Visual Computing, École des Ponts ParisTech, France.

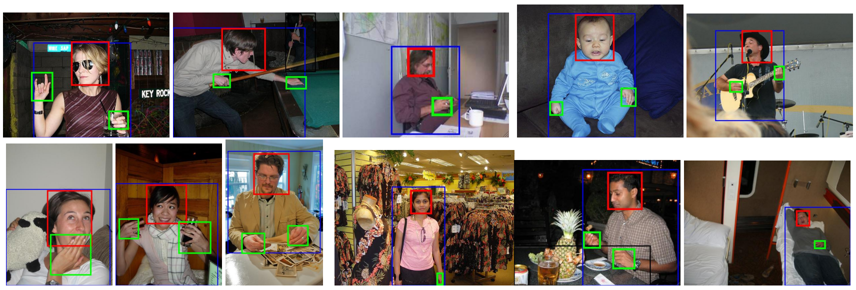


Fig. 1. Sample high ranked results of person layout detection task for the VOC 2010 test dataset. The blue rectangle represents the provided bounding box of the person, green rectangles are the detected hands and red rectangle is the detected head respectively. Our method yields the best results despite high variation in pose and occlusion.

better than misclassifying it as a banana). We refer to the problem of multi-class classification with a structured loss defined by a taxonomy as *taxonomic multi-class prediction*. A taxonomy defines a limited number of losses, which are logarithmic in the number of classes [1].

Person layout is another example of a multi-class problem. The person layout problem of the PASCAL VOC challenge [2] involves predicting the bounding box for several parts of a person (head, hands and feet). It is a very challenging task with evaluation over real-world images comprising a variety of different viewpoints and highly varied human layout configurations. The success of the layout prediction is judged by the correct prediction for the presence/absence of the parts and also by their correct localization. The natural loss underlying this problem can be computed from the number of incorrectly predicted parts.

The output space of both these problems is structured and can be solved by the structured output support vector machine (SOSVM) [3, 4], which generalizes the support vector machine to the case of complex or interdependent output spaces. SOSVM has been successfully applied to core computer vision tasks such as stereo vision [5], object detection [6], and segmentation [7]. Furthermore, both multi-class taxonomic prediction and person layout exhibit an inherent order among the different instances. For any class in the tree multi-class taxonomic prediction the class taxonomy defines the ordering of predicted classes according to the distance in the tree. In the person layout problem, instances having more correctly predicted parts are ranked higher than those with less correctly predicted parts. As such, it is clear that these problems could benefit from the use of structured output *ranking*, which has been shown to improve over the SOSVM when predictions involve ranking results [8–10].

Although structured output ranking often performs better than the SOSVM for problems with certain structured losses, it has previously not been feasible to train the ranking objective with all pairs of training samples when the number of samples is large [11]. The number of constraints in an SVM or SOSVM is linear in the number of training samples, whilst structured output ranking is quadratic. As standard computer vision datasets contain thousands of samples, complexity quadratic in the number of samples is typically infeasible. In all the efficient formulations of structured output ranking proposed so far, the constraints are defined either between ground-truth and one (most) incorrect prediction per training image [8] or between predictions above and below a certain threshold value [9, 10]. If these formulations are generalised for all pairs of predictions (our case), it would result in an $\mathcal{O}(n^2)$ algorithm. In this paper we show that training of the structured output ranking objective can be performed with linear complexity if the structured output loss has a small number of discrete values. We demonstrate the applicability of this linear time algorithm to the two quite disparate tasks: taxonomic multi-class prediction (Section 4) and person layout (Section 5).

We show that learning with structured output ranking consistently surpasses the predictive performance of SVM, SOSVM, ordinal regression and other related methods.

2 Structured Output Ranking – Review

For a given training set, $(x_1, y_1), ..., (x_n, y_n)$, where x_i is the input and y_i is the output, structured output prediction [3] learns a compatibility function $f(x_i, y_i) = \langle w, \phi(x_i, y_i) \rangle$ that assigns a scalar value indicating the fitness of the structured prediction y_i to x_i. To do so, a combined feature representation ($\phi(x_i, y_i)$) is used and the training problem involves the learning of a classifying hyperplane (w) for this joint feature map. For computer vision problems, $x_i \in \mathcal{R}^d$ is the i^{th} image represented as a d-dimensional feature vector, and $y_i \in \mathcal{Y}$ is a sample in a structured output space. The domain of $\mathcal{Y}$ is application specific. For multi-class prediction, $\mathcal{Y} \equiv \{1, ..., c\}$, where c is the number of class labels. For human layout, $\mathcal{Y} \equiv \mathcal{R}^{4r}$, where r is the number of body parts each represented by four co-ordinates of its bounding box.

Ordinal regression. In ordinal regression [12] (which is not a structured output), the output y_i is a scalar value indicating the ordering (rank) of x_i. Thus, for this problem $y_i \in \{1, ..., R\}$, so that the values $1, ..., R$ are related on an ordinal scale. R is the number of rank values. The goal is to learn a ranking function $h(x) = w^T x$, such that higher ranked pairs are assigned higher score, i.e., $h(x_i) > h(x_j) \iff y_i > y_j$. This results in the following optimization problem:

$$\begin{aligned} &\min_{w,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i,j}\xi_{ij} \qquad s.t., \xi_{ij} \geq 0, \\ &w^T x_i - w^T x_j \geq 1 - \xi_{ij} \end{aligned} \tag{1}$$

where (i, j) are the ordered indices of training samples, such that $y_i > y_j$. This formulation finds a solution that minimizes the number of training examples that are swapped w.r.t. their desired order.

Structured output ranking. In structured output *ranking*, the goal is to learn a compatibility function $f(x_i, y_i) = w^T\phi(x_i, y_i)$, such that input-output pairs with lower loss (e.g. fewer mispredicted parts in the case of human layout) are assigned higher compatibility score. To do so, a loss value $\Delta(y_i)$ is defined for every input-output pair (x_i, y_i) which represents the loss associated with the prediction y_i in the image x_i. In this case (x_i, y_i) now correspond to a training set including incorrect predictions y_i for the x_i as well as correct predictions. We will write Δ_i in place of $\Delta(y_i)$ in the sequel. The compatibility function f is learned such that it holds, $f(x_i, y_i) > f(x_j, y_j) \iff \Delta_i < \Delta_j$. Structured output ranking generalises ordinal regression to structured space and does so by modifying the hinge loss paid for misordered pairs. The objective function is modified as follows such that it pays a loss proportional to the difference in losses for misordering:

$$\min_{w,\xi} \; \frac{1}{2}\|w\|^2 + C \sum_{(i,j)\in\mathcal{P}} \xi_{ij} \qquad s.t., \xi_{ij} \geq 0, \tag{2}$$

$$w^T\phi(x_i, y_i) - w^T\phi(x_j, y_j) \geq 1 - \frac{\xi_{ij}}{\Delta_j - \Delta_i}$$

where $\mathcal{P}$ denotes the set of ordered training sample pairs such that the structured output loss, of the sample Δ_i, of sample i is less than the loss, Δ_j, of sample j.

This form (2) of the objective function is referred as slack rescaling. Another common form, margin rescaling, uses the difference in the losses to change the margin directly rather than scaling the slack variable. Due to space restrictions, in the paper we will only consider the case of slack rescaling.

3 Linear Time Constraint Generation

The above formulation has $m \in \mathcal{O}(n^2)$ number of constraints and slack variables, where n is the number of training samples. Although the objective function can contain a number of constraints quadratic in the number of structured output predictions, many vision applications have loss functions that have a fixed number of discrete values, L. For instance, the VOC detection overlap score [2] is a continuous value lying between 0 and 1. However, if it is rounded to tenths of a decimal then we get 11 discrete values. A limited number of discrete loss values also arise in taxonomic multi-class prediction (Section 4) and part based models with thresholded loss (Section 5). Having a small number of loss values enables us to adapt the cutting plane strategy of Joachims [13], achieving a linear time complexity to optimize over a quadratic number of constraints. It is to be noted that in [13], algorithms are given for SVM classification and ordinal regression, which are very different from the problem we are addressing.

We propose an equivalent 1-slack formulation [13] of structured output ranking that uses a single slack variable $\xi = \sum_{(i,j)\in\mathcal{P}} \xi_{ij}$ resulting in the following cutting plane optimization problem:

$$\min_{w,\xi} \; \frac{1}{2}\|w\|^2 + C\xi, \qquad s.t., \xi \geq 0, \tag{3}$$

$$\frac{1}{m}\sum_{(i,j)\in\mathcal{P}} c_{ij}(w^T\phi(x_i, y_i) - w^T\phi(x_j, y_j)) \geq \frac{1}{m}\sum_{(i,j)\in\mathcal{P}} c_{ij} - \frac{\xi}{\Delta_j - \Delta_i}$$

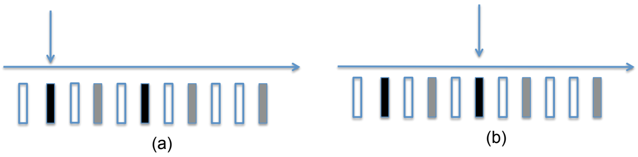


Fig. 2. Linear time constraint generation method. For a particular w the data-points with training pairs (x_i, y_i) are ordered in decreasing compatibility scores $w^T\phi(x_i, y_i)$ (shown by the horizontal arrow sign). The vertical arrow represents the current position of the scan through the data. Gray bars are all the data-points with training pairs having the given loss value l. Black bars are the data-points with training pairs which have loss $l' > l$. For simplicity of the explanation, the margin is considered to be zero. In this case, a constraint is violated when a black bar is placed earlier than a gray bar in the sorted list. In (a), the first black bar will violate constraints with the three subsequent gray bars and in (b), the second black bar will violate constraints with the two remaining gray bars. If an account is maintained of all the subsequent gray bars in the list, then all the violated constraints (i, j) having $\Delta_i = l$ can be obtained in a single scan through the data.

Here $c_{ij} \in \{0, 1\}$ is the indicator variable stating whether the constraint between samples i and j has been included in the summation. The indicator c_{ij} is 1 if the corresponding constraint in (2) is violated and zero otherwise, i.e., $c_{ij} = |(\Delta_i < \Delta_j) \wedge ((w^T\phi(x_i, y_i)) - (w^T\phi(x_j, y_j)) < 1)|$. Thus, a constraint is violated when the training samples are scored such that the difference in the scores of lower loss sample and higher loss sample is less than the margin. While the above formulation (3) has 2^n constraints one for each possible value of c_{ij}, it has only one slack variable ξ shared across all constraints. Each constraint in this formulation correspond to the sum of a sub-set of constraints from formulation (2) selected by c_{ij}.

Cutting plane optimization of Equation (3) consists of alternating between optimizing the objective with a fixed set of constraints, and finding violated constraints of the current function estimate. The core of the learning algorithm is to determine, for a given w, the most violated constraints.

For a given loss value l, all the violated constraints for pairs (i, j) such that $\Delta_i = l$ can be obtained in linear time using a strategy analogous to the one proposed in [13] for ordinal regression. To do so, first the training samples are sorted in terms of decreasing compatibility scores $(w^T\phi(x_i, y_i))$. If a sample j with $\Delta_j > l$ is scored such that difference between the scores of samples i and j is less than one (i.e., the margin is violated), it will also violate all the subsequent samples, i' with $\Delta_{i'} = l$ in the sorted list (i.e., having $w^T\phi(x_{i'}, y_{i'}) < 1 + w^T\phi(x_j, y_j)$). Thus by bookkeeping all the samples having loss value l, all the violated constraints for pairs (i, j) having $\Delta_i = l$ can be found in one pass through the training data. Figure 2 illustrates the method. If the number of possible loss values is small, then this gives a linear time solution for generating all the violated constraints (by going through each of the loss values in this manner). The complete learning method is summa-

Algorithm 1 1-slack optimization for structured output ranking with slack rescaling.

```
 1: Input: $S = ((\phi(x_1, y_1), \Delta_1), ..., (\phi(x_n, y_n), \Delta_n)), C, \epsilon$
 2: $L = (\Delta_1, \Delta_2, ..., \Delta_n)$
 3: sort $L$ in decreasing order
 4: $W \leftarrow \emptyset$
 5: repeat
 6:   $(w, \xi) \leftarrow argmin_{w,\xi \geq 0} \frac{1}{2} w^T w + C\xi$
      $s.t. \forall (c^+, c^-) \in W : \frac{1}{m} w^T \sum_{i=1}^n (c_i^+ - c_i^-)\phi(x_i, y_i) \geq \frac{1}{2m} \sum_{i=1}^n (c_i^+ + c_i^-) - \xi$
 7:   sort $S$ by decreasing $w^T \phi(x_i, y_i)$
 8:   $c^+ \leftarrow 0; c^- \leftarrow 0$
 9:   for $l = l_2, ..., l_{|L|}$ do
10:     $n_l \leftarrow$ number of examples with $\Delta_i = l$
11:     $i \leftarrow 1; j \leftarrow 1; a \leftarrow 0; b \leftarrow 0;\ d \leftarrow 0$
12:     while $i \leq n$ do
13:       if $\Delta_i == l$ then
14:         while $(j \leq n) \wedge (w^T \phi(x_i, y_i) - w^T \phi(x_j, y_j)) < 1$ do
15:           if $\Delta_j > l$ then
16:             $b{+}{+};\ d \leftarrow d + \Delta_j; c_j^- \leftarrow c_j^- + (n_l - a)(\Delta_j - \Delta_i)$
17:           end if
18:           $j{+}{+}$
19:         end while
20:         $a{+}{+}; c_i^+ \leftarrow c_i^+ + d - b\Delta_i$
21:       end if
22:       $i{+}{+}$
23:     end while
24:   end for
25:   $W \leftarrow W \cup \{(c^+, c^-)\}$
26: until $\frac{1}{2m} \sum_{i=1}^n (c_i^+ + c_i^-) - \frac{1}{m} \sum_{i=1}^n (c_i^+ - c_i^-)(w^T \phi(x_i, y_i)) \leq \xi + \epsilon$
27: return $(w, \xi)$
```

rized in Algorithm 1. Source code of a reference implementation is available from our website at: `http://www.robots.ox.ac.uk/~vgg/software/struct_rank/`.

In the algorithm, we have used variables c_i^+ and c_i^- which are the weighted counts of violations in which the ith sample occurs with positive sign (i.e. $c_{ij} = 1$) and negative sign (i.e. $c_{ji} = 1$) respectively. In slack rescaling, these weights are related to c_{ij} by $\frac{1}{m} \sum_{(i,j) \in \mathcal{P}} (\Delta_j - \Delta_i) c_{ij} = \frac{1}{2m} \sum_{i=1}^n (c_i^+ + c_i^-)$. In each iteration, the algorithm computes the optimum over the current working set W (line 6). In lines 9–24, it finds all the violated constraints and adds them to the current working set W (line 25). Unlike ordinal regression, the variables c_i^+ and c_i^- are scaled by the loss values, which results in the scaling of hinge loss of the objective function with the difference in loss values of samples. As long as the number of loss values, $|L|$, is independent of the number of samples, computation of the outer loop (lines 9–24) is therefore also linear in the number of samples. The linear time algorithm for ordinal regression proposed in [13] is a special case of our method. Plugging $\phi(x_i, y_i) = x_i$ and replacing $\Delta_j - \Delta_i$ with 1, our method reduces to ordinal regression.

4 Taxonomic Multi-class Prediction

Class hierarchies (otherwise known as taxonomies) are especially important in computer vision when scaling classification and detection to large numbers of categories [14]. In this work, class hierarchies are learned in a ranking setting.

This is done by using the taxonomic loss as the Δ value in the structured output ranking objective. We define the taxonomic loss as the minimum path length between the two classes in the taxonomic graph, resulting in a number of loss values logarithmic in the number of classes to be predicted [15]. A class nearer to the reference class in the taxonomic graph is ranked higher. For example, if the taxonomy is {{animal, {horse, cow}}, {vehicle, {bus, motorbike}}}, then the loss for misclassifying a 'cow' with 'horse' will be 2 whereas the loss for mis-classification between 'cow' and 'motorbike' will be 4. Thus, the number of possible loss values is $\mathcal{O}(\log c)$, where c is the number of class labels. It should be noted that usually $c \ll n$,where n is the number of training samples.

The joint feature map, ϕ used in this work, is the standard one used in multi-class and taxonomic prediction [15, 16]. This is given by $\phi(x_i, y_i) = \lambda(y_i) \otimes x_i$. $\lambda(y_i) \in \mathcal{R}^c$ is the class attribute vector, which is defined as: $\lambda_j(y_i) = 1$, if $j = y_i$, zero otherwise. $x_i \in \mathcal{R}^d$ is the i^{th} image represented as a d-dimensional feature vector. Here $\otimes$ denotes a Kronecker product, thus $\phi(x_i, y_i) \in \mathcal{R}^{d \cdot c}$. The ordinal constraints of (2) therefore enforce that the scores of the classes are ordered in proportion to their distance in the taxonomic tree to a ground truth class.

Evaluation measure. The performance is evaluated using the cumulative taxonomic loss [15]. This is computed by accumulating the taxonomic loss over the top scoring t classes for a given image, where $t \in [1..C]$. The results are reported (Section 6) as the mean cumulative taxonomic loss which is obtained by averaging the cumulative taxonomic loss for different ranks over all the test images. The ImageNet challenge [17] also uses a hierarchical cost as an evaluation criteria.

We test how well structured ranking performs compared to other methods for the following two datasets:

Indoor scene database. The indoor scene database [18] consists of 15620 images for 67 different indoor categories. The categories are further grouped into 5 scene groups which defines a two level taxonomy. The dataset is partitioned into the training set and the test set by choosing 80 and 20 images from each of the classes, respectively.

PASCAL VOC 2007 classification dataset. The PASCAL VOC 2007 classification dataset [2] is comprised of 5011 images in the combined training and validation set and 4952 images in the test set. Each image may contain multiple objects and the images are annotated for 20 different object categories. A taxonomy is defined with these 20 classes by the organizers [2], which we use for our experiments.

4.1 Implementation Details

Image descriptors. SIFT descriptors [19] are extracted with a spatial stride of 5 pixels, and at four scales, defined by setting the width of the SIFT spatial bins to 4, 6, 8 and 10 pixels respectively. For the indoor scene database experiments, the features are quantised into a visual vocabulary of size 1000. A 2-level spatial pyramid model [20] is obtained by dividing the image in 1×1 and 2×2 grids, for a total of 5 regions. For the experiments on the VOC 2007 dataset, the visual vocabulary used is of size 10000, and the quantized descriptors are encoded using locality-constraint linear encoding (LLC) [21]. The spatial regions are obtained by dividing the image in 1×1, 3×1 and 2×2 grids, for a total of 8 regions.

Learning methods. SVM and SVM-rank classifiers are trained by minimizing the 0/1 loss and are learned in a 1 vs. all fashion. SOSVM and structured ranking SVM algorithms minimize the taxonomic loss and are modelled using the joint feature map.

5 Person Layout

In the person layout competition of the VOC challenge [2], a bounding box (refered to hereafter as a human ROI) is provided for each 'person' object in a test image. The job is then to predict the presence or absence of parts (head/hands/feet), and the bounding boxes of those parts. The prediction for a person layout should be output with an associated real-valued confidence of the layout. This confidence score is then used to compute the precision-recall (PR) curve and AP (area under the PR curve) for each of the parts. The mean AP across all the parts is considered as the AP for the person. A submission is evaluated by the AP for individual parts and also by its summary for the person. Therefore, in order to win the competition, it is imperative to perform well on each of the individual parts categories.

Current layout detection methods tend to model the human body as a pictorial structure, and to predict the pose/layout of the person by maximising the posterior of the joint configuration of the body parts [22, 23]. However, these methods suffer from a common curse of pictorial structures, the inability to model occlusion of parts and the over-counting of confidence from a given pixel location. In the experiments, we show that unmodified pictorial structure models are not suited to this setting.

Another approach to solve the person layout problem could be to apply individual part detectors and then combine their outputs in some fashion. There are two caveats here: (i) the task needs the confidence score for the whole layout, not the individual parts; (ii) detection of parts is performed independently, which may lead to sub-optimal performance unless some kind of co-occurrence information is introduced into the framework.

The method we implement here proceeds in two stages, (i) part candidates are generated using individual detectors, (ii) candidates for individual parts are combined and the joint output space is optimized and ranked using the structured output ranking approach. We note that feet are not detected in the current experiments as the training set size is not sufficient to reliably estimate the wide range of appearance in the test set (The AP of the best performing foot detection method for the PASCAL VOC 2010 competition was 1.2%).

Experiments are performed using the PASCAL VOC 2011 [24] person layout dataset. There are 609 images in both the training and validation sets having 850 human ROIs. The results are reported as the mean AP and standard deviation (Section 6).

5.1 Part candidates generation

The candidates for heads and hands are generated using individual part detectors. Head candidates are generated using a part-based detector of the human head [25]. For the generation of hand candidates, the hand detector available from [26] is used. The output of both the detectors is a bounding box around the part (refered to as ROI hereafter) with a confidence value. The candidates for the parts so obtained are rescored using separate linear SVM classifiers. This is done to include local image information such as positional and scale attributes of the candidates, relative to the human ROI in the image.

The linear SVM classifier for head candidates is trained over a feature vector (Φ_{head}) formed by concatenating: (i) confidence score from the detector, (ii) relative size of head and human ROIs, (iii) fraction of area of head ROI in human ROI. The intuition behind using these features is to suppress any bounding box which either has an abnormal height with respect to the size of the image, or which does not fully lie within the given bounding box of the person. For every human ROI the top scoring head candidate is returned as the detected head.

The linear classifier for the hand is trained in a similar fashion. The feature vector (Φ_{hand}) consists of: (i) confidence score of the detector, (ii) position within human ROI, (iii) relative position of hand and head ROIs within human ROI, (iv) relative size of hand and head ROIs, (v) overlap score of head and hand ROIs, (vi) fraction of area of hand ROI in human ROI. A head can be detected more reliably in an image compared to a hand. Therefore, the information about the head position is utilized for the detection of hands. The scores of hand candidates are thresholded to give 75% recall on the training set. For every person a maximum of top two scoring hand bounding boxes are returned as the detected hands (zero or one hand may be returned if no image regions score above the threshold).

5.2 Joint learning using structured ranking

We obtain confidence score for the person layout from a ranking function which is learned by jointly optimizing the output spaces of the two parts using structured output ranking. We also compare our results with naïve techniques for combining the confidences of different parts. However, they do not optimise the AP for all parts jointly and as we will show tend to benefit one of the parts at the expense of others.

We define the loss value (Δ) for a human layout as 1 - precision. This defines a limited number of loss values corresponding to the fraction of hypothesized part detections that are incorrect. For example, if all parts are hypothesized to be present, then the set of possible loss values is: $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. The numbers in the bracket correspond to the losses when {all, two, one, none} of the hypothesized body parts are correct. The ranking objective function is trained on the

		Indoor			**VOC 2007**		
Method	**Obj Loss**	**1**	**5**	**10**	**1**	**5**	**10**
SVM	0/1	2.75	16.18	33.51	2.83	17.09	39.84
SVM-struct	tax	3.71	18.50	36.97	3.98	20.71	45.19
SVM-rank	0/1	3.14	15.93	32.11	3.02	17.04	39.29
Struct rank	tax	**2.62**	**14.75**	**30.29**	**2.54**	**15.07**	**36.07**
Min loss		0.00	8.00	18.00	0.00	11.70	33.35

Table 1. Mean cumulative taxonomic loss at different ranks. The cumulative loss is computed as explained in Section 4. 'Obj loss' is the loss that is minimized by the respective learning method, where 'tax' stands for the taxonomic loss. 'Min loss' is the theoretically possible minimum taxonomic loss which is obtained by optimally ordering the class labels. The minimum loss is zero for the top result as the taxonomic loss is zero for correctly labeling an image. For the top two ranked classes, the minimum cumulative loss is the lowest taxonomic loss of a second category. This measure differentiates methods that correctly order multiple predictions by their taxonomic loss. The mean cumulative taxonomic loss for the structured output ranking method is lowest among all the learning methods.

feature vector formed by concatenating the features used earlier for rescoring of parts (i.e., by concatenating the feature vectors of head and all the hands). Thus, $\phi(x_i, y_i) = [\Phi_{head}\, (\Phi_{hand})^h]$, where h is the number of hands in the i^{th} image, therefore, $\phi(x_i, y_i) \in \mathcal{R}^{3+6h}$.

6 Experimental Results

In this section we show that using structured ranking SVM a significant improvement is achieved for both the applications. Due to the space constraints, results for slack rescaled versions of structured SVMs are reported. We also experimented with margin rescaled version and found that it performs slightly worse than slack rescaled version, but the difference was negligible.

6.1 Taxonomic Multi-class Prediction

In the following experiments, we compare our approach with binary SVM, SOSVM (SVM-struct) and ordinal regression (SVM-rank). For experiments on the indoor scene database, we train our model using the training set and test it on the test set. For the VOC 2007 dataset, training is done using the training-validation set and testing on the test set following the competition protocol. A validation step was employed to set the C parameter. For all formulations linear kernels were used.

Table 1 gives the numbers for cumulative loss for the top 1, 5 and 10 ranks for the different methods. It can be seen that the cumulative taxonomic loss decreases for ranking algorithms and it reduces even further by using structured ranking algorithms. Some qualitative results are shown in Figure 3.

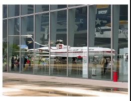
	Indoor			**VOC 2007**		
						
SVM	airport, restaurant, mall	bath, elevator, cloister	deli, museum, studio	person, chair, tv	train, cow, bus	car, bus, tv
SVM struct	salon, audi, bowling	deli, florist, mall	church, florist, studio	cat, car, bus	sheep, horse, cat	chair, sofa, sheep
SVM rank	video, grocery, bookstore	cellar, corridor, childroom	locker, pool, prison	sofa, chair, table	train, bus, cycle	table, tv, plant
Struct rank	mall, grocery, bookstore	elevator, waiting, pool	museum, locker, church	chair, sofa, plant	train, plane, boat	person, plane, tv

Fig. 3. Sample results for the indoor scene and the VOC 2007 classification test datasets. The top three labels (in order) predicted by each method are shown for several test images. More semantically meaningful labels are retrieved by structured ranking than the other methods.

The performance is evaluated over taxonomic loss, therefore, an objective function minimizing taxonomic loss is supposed to give the optimal performance. However, SOSVM (labelled 'SVM-struct') performs *worse* on these data sets than a binary SVM even when it optimizes taxonomic loss. A SVM optimizing ordinal regression with binary loss (SVM-rank) tends to perform worse than a binary SVM when few labels are returned, but improves over the binary SVM when more labels are returned. However, structured output ranking dominates the performance of all other learning variants. This shows that structured output ranking is the optimal way of optimizing taxonomic loss for the given scenario. The mean cumulative taxonomic loss for the structured output ranking method is closer to the minimum loss for the VOC 2007 dataset than the indoor scene database. This may be because the VOC 2007 dataset has a deeper taxonomy.

6.2 Person Layout

For all the following experiments, 5-fold cross-validation is performed using the training and validation set of the PASCAL VOC 2011 [24] person layout dataset.

Part candidates generation. After rescoring the head and hand candidates using linear SVM classifiers, AP for the head part increases from (68.59±3.14) to (81.04±1.7) and for hands from (23.97±2.24) to (28.32±2.12). It shows that the modeling of dependencies between parts improves detection performance. In general, this can be thought of as a (directed) graphical model that encodes the dependencies between parts. On the same images Yang et al.'s method [23] gives (37.71±1.73) AP for head and (0.27±0.02) AP for hand. Their system was designed using a stickman for person layout (i.e., one line segment indicating

Ordering by	**Head AP%**	**Hand AP%**	**Mean**
Head score	81.04±1.77	20.98±2.21	51.01±1.23
Max-hand score	74.95±3.16	26.16±2.49	50.55±1.55
Mean-hand score	75.61±2.33	28.32±2.12	51.96±1.15
Mean head, hands	79.77±2.20	22.90±2.27	51.33±1.10
Max head, hands	79.49±2.00	21.53±2.30	50.51±1.25

Table 2. AP for different scoring methods. The confidence score for the person layout prediction is computed by combining parts scores in different ways. The last column is the mean of head and hand APs, which is the metric to be maximized. Experiments were performed on train-val set of the VOC 2011 layout dataset.

Method	**Head**	**Hand**	**Mean**
SVM linear	73.92±3.15	20.29±1.76	47.1±1.87
Rank linear	79.32±2.77	27.88±1.75	53.6±1.28
Rank RBF	79.55±2.88	28.22±2.25	**53.9±1.29**

(a)

Method	**Head**	**Hand**	**Mean**
Our Method	72.85	26.7	**49.8**
BCNPCL	74.4	3.3	38.8
OXFORD	52.7	10.4	31.5

(b)

Table 3. (a) AP scores resulting from different learning techniques. The last two rows are different variants of the proposed method. They differ only slightly, but improve substantially over the SVM. The dataset used for the experiments was train-val set of the VOC 2011 layout dataset. **(b) AP for the VOC 2010 person layout test dataset.** We train our method on the train-val portion of the VOC 2010 layout dataset. The evaluation was computed on the competition server. The results for the other methods are as reported on the competition website [27]. Our result for hand detection is even better than [26], which reports AP of 23.18 for the same dataset.

location, size and orientation for each part). To evaluate their detections using the PASCAL criteria, we estimate the bounding boxes from the stickman. The height is given by length of the stickman line segments and the width is fixed by cross-validation. The hand bounding box is detected by extrapolating along the orientation of the arm from the detected wrist end-point. The poor performance of Yang et al.'s method shows that unmodified pictorial structure approaches do not successfully solve the current problem.

Joint learning using structured ranking. For the experiments, both linear and non-linear kernels were explored. The C parameter and the σ parameter of the Gaussian RBF kernel are optimized by a validation step.

Table 2 shows the inherent tradeoff between the confidences of various parts. While naïve combination schemes can be employed to compute a joint detection score based on the individual part scores, such a strategy will likely choose a suboptimal score. By directly optimizing a loss computed over the combination of parts, the structured output ranking objective can better balance the performance of each part detection. Table 3(a) shows the improved performance when an appropriate joint scoring objective is applied. We see that the ranking objectives (c.f. Section 2) perform substantially better than a binary SVM (first row) and improves the AP for both head and hands.

Comparison with PASCAL VOC 2010 results. Table 3(b) compares our results with the best performing results in the PASCAL VOC 2010 layout detection competition. We achieve a substantial improvement over the wining entries using our method. For the comparison, we trained the model on train-val set of the VOC 2010 layout dataset. Figure 1 shows some sample results for person layout.

Thus, the person layout experiments show both the effects of the prediction architecture and feature design, as well as the benefits of using a structured output ranking formulation.

Time analysis. Run-time is measured for training on train-val set of the PASCAL VOC 2011 layout dataset. Our method takes on average about $1/80^{th}$ of a second whereas a normal $\mathcal{O}(n^2)$ implementation takes around 0.5 seconds (i.e., takes 40 times more time) per cutting plane iteration for a linear kernel training.

7 Conclusions

The experiments in Section 6 lead to several broad conclusions. There is a consistent improvement when using structured output ranking over structured output SVMs and ordinal regression. This indicates that structured output ranking is able to encapsulate the benefits of both approaches, leading to better overall performance. Furthermore, the training of structured output ranking is no more computationally expensive than that of binary SVMs or ordinal regression (c.f. Section 3). Using our method we also achieve state-of-the art results for the PASCAL VOC human layout problem.

The efficient linear time training algorithm for structured output ranking can be used whenever the number of loss values is small and independent of the number of training samples. This is true for a large variety of practical problems. In this work, we have demonstrated this to be the case for taxonomic multi-class prediction and person layout. As such diverse applications as scene layout, object layout, multi-class and multi-label prediction are characterized by small numbers of loss values, and also methods that rank learning from continuous loss values such as [28] can simply be discretized into a small number of losses, we expect that the methods proposed here will find wide application across learning based computer vision.

Acknowledgements. We thank Ken Chateld and Mayank Juneja for providing us with SIFT descriptors for the experiments. We are grateful for financial support from ERC grant VisRec no. 228180, ERC grant no. 259112, ONR MURI N00014-07-1-0182 and PASCAL2 Network of Excellence.

References

1. Bengio, S., Weston, J., Grangier, D.: Label embedding trees for large multi-class tasks. In: NIPS. (2010)

2. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC) challenge. IJCV (2010)
3. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: Proc. ICML. (2004)
4. Joachims, T., Finley, T., Yu, C.N.J.: Cutting-plane training of structural SVMs. Machine Learning (2009)
5. Li, Y., Huttenlocher, D.P.: Learning for stereo vision using the structured support vector machine. In: Proc. CVPR. (2008)
6. Blaschko, M.B., Lampert, C.H.: Learning to localize objects with structured output regression. In: Proc. ECCV. (2008)
7. Szummer, M., Kohli, P., Hoeim, D.: Learning CRFs using graph cuts. In: Proc. ECCV. (2008)
8. Blaschko, M.B., Vedaldi, A., Zisserman, A.: Simultaneous object detection and ranking with weak supervision. In: NIPS. (2010)
9. Rahtu, E., Kannala, J., Blaschko, M.B.: Learning a category independent object detection cascade. In: Proc. ICCV. (2011)
10. Zhang, Z., Warrell, J., Torr, P.H.S.: Proposal generation for object detection using cascaded ranking SVMs. In: Proc. CVPR. (2011)
11. Huang, J.C., Frey, B.J.: Structured ranking learning using cumulative distribution networks. In: NIPS. (2008)
12. Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression. In: Advances in Large Margin Classifiers. (2000)
13. Joachims, T.: Training linear SVMs in linear time. In: KDD. (2006)
14. Deng, J., Berg, A.C., Li, K.and Li, F.F.: What does classifying more than 10, 000 image categories tell us? In: Proc. ECCV. (2010)
15. Binder, A., Müller, K.R., Kawanabe, M.: On taxonomies for multi-class image categorization. IJCV (2011)
16. Cai, L., Hofmann, T.: Exploiting known taxonomies in learning overlapping concepts. In: IJCAI. (2007)
17. Imagenet: http://www.image-net.org/
18. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: Proc. CVPR. (2009)
19. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
20. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: Proc. CVPR. (2006)
21. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. Proc. CVPR (2010)
22. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. In: Proc. BMVC. (2009)
23. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: Proc. CVPR. (2011)
24. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2011 (VOC2011). http://www.pascal-network.org/challenges/VOC/voc2011/ (2011)
25. Marin-Jimenez, M., Zisserman, A., Ferrari, V.: Heres looking at you, kid. detecting people looking at each other in videos. In: Proc. BMVC. (2011)
26. Mittal, A., Zisserman, A., Torr, P.H.S.: Hand detection using multiple proposals. In: Proc. BMVC. (2011)
27. VOC2010-Results: http://pascallin.ecs.soton.ac.uk/challenges/voc/voc2010/results/
28. Li, F., Carreira, J., Sminchisescu, C.: Object recognition as ranking holistic figure-ground hypotheses. In: Proc. CVPR. (2010)