Max Margin AND/OR Graph Learning for Parsing the Human Body

Long (Leo) Zhu
Department of Statistics
University of California, Los Angeles
lzhu@stat.ucla.edu

Yuanhao Chen
University of Science and Technology of China
yhchen4@ustc.edu

Yifei Lu
Shanghai Jiao Tong University
klux@sjtu.edu.cn

Chenxi Lin
Microsoft Research Asia
chenxil@microsoft.com

Alan Yuille
Department of Statistics, Psychology and Computer Science
University of California, Los Angeles
yuille@stat.ucla.edu

Abstract

We present a novel structure learning method, Max Margin AND/OR Graph (MM-AOG), for parsing the human body into parts and recovering their poses. Our method represents the human body and its parts by an AND/OR graph, which is a multi-level mixture of Markov Random Fields (MRFs). Max-margin learning, which is a generalization of the training algorithm for support vector machines (SVMs), is used to learn the parameters of the AND/OR graph model discriminatively. There are four advantages from this combination of AND/OR graphs and max-margin learning. Firstly, the AND/OR graph allows us to handle enormous articulated poses with a compact graphical model. Secondly, max-margin learning has more discriminative power than the traditional maximum likelihood approach. Thirdly, the parameters of the AND/OR graph model are optimized globally. In particular, the weights of the appearance model for individual nodes and the relative importance of spatial relationships between nodes are learnt simultaneously. Finally, the kernel trick can be used to handle high dimensional features and to enable complex similarity measure of shapes. We perform comparison experiments on the baseball datasets, showing significant improvements over state of the art methods.

1. Introduction

Parsing the human body (i.e. pose estimation of body parts) in static image has received a lot of attention. Such problems arise in many applications including human action analysis, human body tracking, and video analysis. But the major difficulties of parsing the human body, which come from the *large appearance variations* (e.g. different clothes) and *enormous number of poses*, are not fully solved. There are three aspects to addressing these problems. Firstly, what representation is capable of modeling the large variation of both shape and appearance? Secondly, how can we learn a probabilistic model defined on this representation? Thirdly, if we have a probabilistic model, how can we perform inference efficiently? (in order to estimate poses for novel images). These three aspects are clearly related to each other. Intuitively, the greater the representational power, the bigger the computational complexity of learning and inference. Most works in the literature, e.g. [18, 9, 3], focus on only one or two aspects, and not on all of them (see section (2.1) for a review of the literature). In particular, the representations used have been comparatively simple. Moreover, attempts to use complex representations tend to specify their parameters by hand and do not learn them from training data.

In this paper, we *represent* the different poses of the human body by the AND/OR graph proposed by Chen *et al* for modeling deformable articulated objects [3]. The advantages of this AND/OR graph (see figure 1) is that it can represent an enormous number of different poses (98 in this paper), enforce (probabilistic) spatial relations on the configuration, and use many image features as input (to address the large appearance variations). Moreover, we *learn* the parameters of this model, which specify the geometry and the appearance, by a novel extension of the max-margin al-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

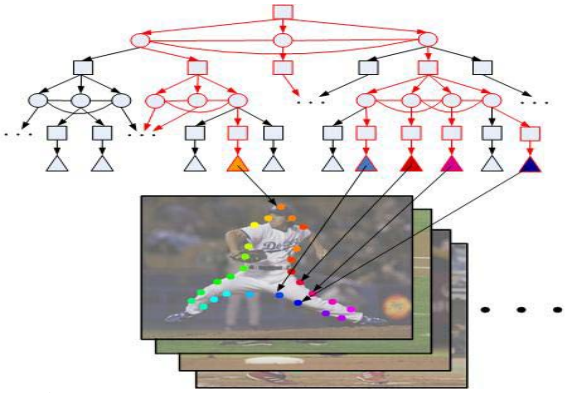

Figure 1. The AND/OR representation allows us to model enormous poses of the object. A parse tree which is a portion of the AND/OR graph represent a specific pose of human body. The nodes and edges with red boundary indicate one parse tree. In this paper, there are 98 poses which can be modeled by the parse trees of the whole AND/OR graph.

gorithm for structure learning [1, 20, 21]. This learning is global in the sense that we learn all the parameters simultaneously (by an algorithm that is guaranteed to find the global minimum) rather than learning subsets of the parameters locally. Max-margin learning has been shown to be more effective than standard maximum likelihood estimation when the overall goal is classification (i.e. into different poses). It also has some technical advantages such as: (i) avoiding the computation of the partition function of the distribution, and (ii) the use of the kernel trick to extend the class of features. To perform *inference*, we use the compositional algorithm described in [3].

Our paper makes contributions to both machine learning and computer vision. The contribution to machine learning is to extend max-margin learning to AND/OR graphs (max-margin has previously been applied to simpler models, see section (2)). The contribution to computer vision is the combination of the AND/OR *representation*, the max-margin *learning*, and compositional *inference* [3] to model human parsing. Moreover, our results, see section (6), show that our approach significantly outperforms the state of the art.

2. Background

2.1. Human Body Parsing

There has been considerable recent interest in human body parsing. Sigal and Black [17] address the occlusion problem by enhancing the ability of appearance modeling. Triggs and his colleagues [15] learn more complex models for individual parts by SVM and combine them by an extra classifier. Mori [9] use super-pixels to reduce the search space and thus speed up the inference. Ren *et al*. [14] present a framework to integrate multiple pairwise constraints between parts. The models of body parts are independently trained. Ramanan [13] propose tree structured CRF to learn a model for parsing human body. Lee and Cohen [8] and Zhang *et al*. [24] used MCMC for inference. In summary, these methods involve representations of limited complexity (i.e. with less varieties of pose than AND/OR graphs). If learning is involved, it is local but not global (i.e. the parameters are not learnt simulateously) [14, 17, 15, 9]. Moreover, the performance evaluation is performed by outputting a list of poses and takes credit if the groundtruth result is in this list [9, 24, 18].

The most related work is by Srinivasan and Shi [18] who introduced a grammar for dealing with the large number of different poses. Their model was manually defined, but they also introduced some learning in a more recent paper [19]. Their results are the state of the art, so we make comparisons to them in section (6).

By contrast, our model uses the AND/OR graph in the form of Chen *et al*. [3] which combines a grammatical component (for generating multiple poses) with a markov random field (MRF) component which represents spatial relationships between components of the model (see [6, 2] for different types of AND/OR graph models). We perform global learning of the model parameters (both geometry and appearance) by max-margin learning. Finally, our inference algorithm outputs a single pose estimate only which, as we show in section (6), is better than any of the results in the list output by Srinivasan and Shi [18] (and their output list is better than that provided by other algorithms [9]).

2.2. Max Margin Structure Learning

The first example of max-margin structure learning was proposed by Altun *et al*. [1] to learn Hidden Markov Models (HMMs) discriminatively. This extended the max margin criterion, used in binary classification [23] and multi-class classification [4], to learning structures where the output can be a sequence of binary vectors (hence an extension of multi-class classification to cases where the number of classes is 2^n, where n is the length of the sequence). We note that there have been highly successful examples in computer vision of max-margin applied to binary classification, see SVM-based face detection [11].

Taskar et al. [20] generalized max margin structure learning to general markov random fields (MRF's), referred to an max margin markov networks(M^3). Taskar et al. [21] also extended this approach to probabilistic context-free grammar (PCFG) for language parsing. But max-margin learning has not yet been extended to learning AND/OR graph models that can be thought of as combining PCFG's with MRF's.

This literature on max-margin structure learning shows that it is highly competitive with conventional maximum likelihood learning methods as used, for example, to learn conditional random fields (CRF's) [7]. In particular, max-margin structure learning avoids the need to estimate the partition function of the probability distribution (which is

major technical difficulty of maximum likelihood estimation). Max-margin structure learning essentially learns the parameters of the model so that the groundtruth states are those with least energy (or highest probability) and states which are close to groundtruth also have low energy (or high probability). See section (5) for details.

3. The AND/OR Graph Representation

3.1. The AND/OR model

The structure of the AND/OR graph is represented by a graph $G = (V, E)$ where V and E denote the set of vertices and edges respectively. The vertex set V contains three types of nodes,"OR" nodes,"AND" nodes and "LEAF" nodes which are depicted in figure (1) by circles, rectangles and triangles respectively. These nodes have attributes including position, scale, and orientation. The edge set E contains vertical edges defining the topological structure and horizontal edges defining spatial constraints on the node attributes. For each node $\nu \in V$, the set of its child nodes is defined by T_ν. Hence $\{T_\nu\}$ denotes all possible vertical edges of the AND/OR graph (the presence of OR nodes means that not all child nodes will appear in a parse, see next subsection). The horizontal edges are defined on triplets (μ, ρ, τ) of the children of AND nodes. The structure of the AND/OR graph is represented by $\{(\nu, T_\nu, (\mu, \rho, \tau))\}$.

The AND/OR graph we use in this paper to represent human pose is shown in figure (2). The top node shows all the 98 possible configurations (i.e. parse trees of the human body). These configurations are obtained by AND-ing sub-configurations such as the torso, the left leg, and the right leg of the body (see circular nodes in the second row). Each of these sub-configurations has different *aspects* as illustrated by the AND nodes (rectangles in the third row). These sub-configurations, in turn, are composed by AND-ing more elementary configurations (see fourth row) which can have different aspects (see fifth row).

3.2. The representational power of the AND/OR Graph Representation

The representational power of AND/OR graph is given by the number of topological configurations of the graph which we call parse trees and which correspond to different poses. Each parse tree corresponds to a specification of which AND nodes are selected by the OR nodes (i.e. each OR node is required to select a unique child). Hence the number of different parse trees is bounded above by W^{K^h}, where K is the maximum number of children of AND nodes (in this paper we restrict $K \leq 4$), W denotes the maximum number of possible children of OR nodes, and h is the number of levels containing OR nodes with more than one child node. The total number of parameters associated with the potential functions, which are defined on the edges of an AND/OR graph, is bounded above by MW^K where M is the number of AND nodes connecting to OR nodes. Hence the AND/OR graph can represent an exponentially large number of articulated poses but with a compact form. This property of the AND/OR graph representation is very desirable for learning because it requires few training images to achieve good generalization. In the experiments reported in this paper we have $M = 35, K = 4, W = 3, h = 4$. There are 98 poses modeled by AND/OR graph.

3.3. The state variables

A configuration (parse tree) of the AND/OR graph is an assignment of state variables $y = \{z_\nu, t_\nu\}$ with $z_\nu = (z^x_\nu, z^y_\nu, z^\theta_\nu, z^s_\nu)$ to each node ν, where (z^x, z^y), z^θ and z^s denote image position, orientation, and scale respectively. The $t = \{t_\nu\}$ variable defines the specific topology of the parse tree, where t_ν denotes the children of node ν. For AND nodes, the set of children is fixed (i.e. not dynamic) and so $t_\nu = T_\nu$. But each OR node must select a unique child node $t_\nu \in T_\nu$ (to enable sub-configurations to switch their appearance, see figure (2)). The input to the graph is the image $x = \{x_\nu\}$ defined on the image lattice (at the lowest level of the graph).

We define $V^{LEAF}(t)$, $V^{AND}(t)$,$V^{OR}(t)$ to be the set of LEAF, AND, and OR nodes which are active for a specific choice of the topology t of a parse tree. These sets can be computed recursively from the root node, see figure (2). The AND nodes in the second row (i.e. the second highest level of the graph) are always activated and so are the OR nodes in the third row. The AND nodes activated in the fourth row, and their OR node children in the fifth row, are determined by the t variables assigned to their parent OR nodes. This process repeats till we reach the lowest level of the graph.

3.4. The potential functions for the AND/OR graph

The conditional distribution on the states and the data is given by:

$$P(y|x; \mathtt{w}) = \frac{1}{Z(x; \mathtt{w})} \exp \langle \mathtt{w}, \Psi(x, y) \rangle . \qquad (1)$$

where x is the input image, y is the parse tree, and $Z(x, \mathtt{w})$ is the partition function. $P(y|x; \mathtt{w})$ is a (conditional) exponential model which is defined by an inner product $\langle \mathtt{w}, \Psi(x, y) \rangle$ between features $\Psi(x, y)$ and model parameters $\mathtt{w}$ (to be learnt). The features $\Psi(x, y)$ are of three types: (i) appearance features $\Psi^D(x, y)$, (ii) horizontal spatial relationship features $\Psi^H(y)$, and (iii) vertical relationship features $\Psi^V(y)$. Note that only the appearance features depend on the data x (the other features are like prior distributions).

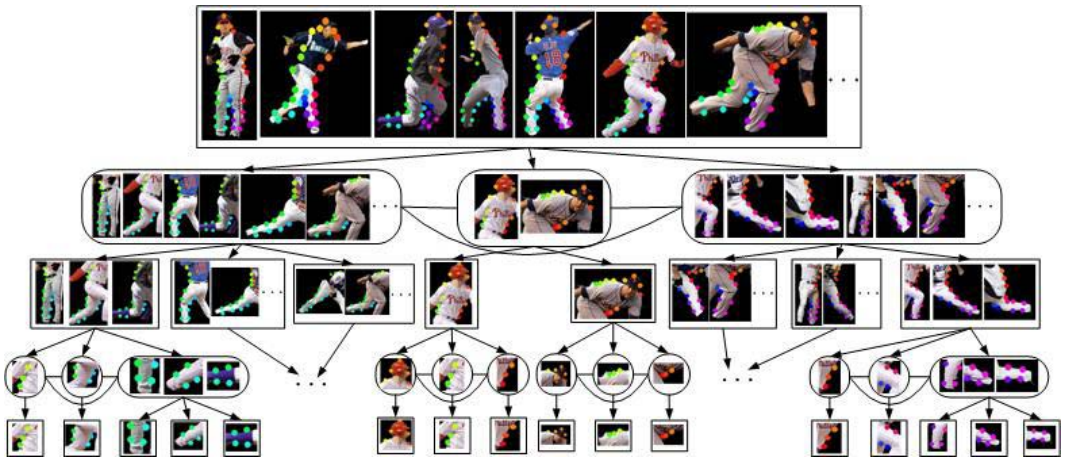

Figure 2. The AND/OR graph is an efficient way to represent different appearances of an object. The graph is built up manually. The bottom level of the graph indicates points along the boundary of human body. The higher levels indicat combinations of elementary configurations. The graph that we used contains eight levels (three lower levels are not depicted here due to lack of space). Color points distinguish different body parts. The arms are not modeled in this paper.

The first type of features $\Psi^D_\nu(x,y), \forall \nu \in V^{LEAF}(t)$ are data dependent and model the appearance of the object. They relate the appearance of the active leaf nodes to properties of the local image. More formally, y in $\Psi^D_\nu(x,y)$ refers to $z_\nu = (z^x, z^y, z^s, z^\theta)$ for the active nodes $\nu \in V^{LEAF}$. $\Psi^D_\nu(x, z_\nu)$ represent the local image features including the grey intensity, gradient, canny edge map, the responses of Gabor filters at different scales and orientations, and related features. We use a total of 101 features of this type (i.e. the vector $\Psi^D(x,y)$ has 101 dimensions). But not all these features will be used (the max-margin learning will typically set some of the parameters w^D to be zero).

The second type of features $\Psi^H(y)$ specify the horizontal relationships (which correspond to geometric constraints at a range of scales). They are defined by $\Psi^H_\nu(y) = g(z_\mu, z_\rho, z_\tau), \forall \nu \in V^{AND}(t)$ where $g(.,.,.)$ is a logarithm of Gaussian distribution defined on the *invariant shape vector* $l(z_\mu, z_\rho, z_\tau)$ [25] constructed from triple child nodes (z_μ, z_ρ, z_τ) of node ν. This shape vector depends only on variables of the triple, such as the internal angles, that are invariant to the translation, rotation, and scaling of the triple. This type of feature is defined over all triples formed by the child nodes of each parent, see figures (2). The parameters of the Gaussians are estimated from the labeled training data (this is local learning, but max-margin will learn their parameters w^H globally).

The third type of features $\Psi^V(y)$ are the vertical components which hold the structure together by relating the state of the parent nodes to the state of its children. $\Psi^V(y)$ is divided into three vertical energy terms denoted by $\Psi^{V,A}(y)$, $\Psi^{V,B}(y)$ and $\Psi^{V,C}(y)$ which refer to type(A), type(B) and type(C) vertical connections respectively.

$\Psi^{V,A}(y)$ specifies the coupling from the AND node to the OR node. This coupling is deterministic – the state of the parent node is determined precisely by the states of the child nodes. This is defined by $\Psi^{V,A}(y) = h(z_\nu, \{z_\mu \text{ s.t.} \mu \in t_\nu\}), \forall \nu \in V^{AND}(t)$, where $h(.,.) = 0$ if the average orientations and positions of the child nodes are equal to the orientation and position of the parent node (i.e. the vertical constraints are "hard"). If they are not consistent, then $h(.,.) = \kappa$, where κ is a small negative number.

$\Psi^{V,B}(y)$ accounts for the probability of the assignments of the connections from OR nodes to AND nodes. We define $\Psi^{V,B}(y) = \lambda_\nu(t_\nu), \forall \nu \in V^{OR}(t)$, where $\lambda_\nu()$ is the potential function which encodes the weights of the assignments determined by t_ν.

The potential function $\Psi^{V,C}(y)$ defines the connection from the lowest AND nodes to the LEAF nodes. This is similar to the definition of $\Psi^{V,A}(y)$, and $\Psi^{V,C}(y)$ is given by $\Psi^{V,C}(y) = h(z_\nu; z_{t_\nu})$ where $h(.,.) = 0$ if the orientation and position of the child (LEAF) node is equal to the orientation and position of the parent (AND) node. If they are not consistent, then $h(.,.) = \kappa$.

4. The Inference/Parsing Algorithm

We use the inference algorithm described in [3] to obtain the best parse tree y^* of an image x by computing $y^* = \arg\max_y \langle \mathrm{w}, \Psi(x,y) \rangle$. This algorithm runs (empirically) in polynomial time in terms of the number of levels of the AND/OR graph (no other algorithm has this level of inference performance on AND/OR graphs). This rapid inference is necessary to make max margin learning practical.

The algorithm has a bottom-up stage which makes proposals for the configuration of the AND/OR graph. This proceeds by combining proposals for sub-configurations to build proposals for larger configuration. For AND nodes, we combine proposals for the child nodes to form a proposal for the parent node. For OR nodes, we enumerate all proposals from all branches without composition. To prevent a combinatorial explosion we prune out weak propos-

Input: $\{MP^1_{\nu 1}\}$. Output:$\{MP^L_{\nu L}\}$. $\oplus$ denotes the operation of combining two proposals.
Loop : $l = 1$ to L, for each node ν at level l

- IF ν is an OR node
  1. Union: $\{MP^l_{\nu,b}\} = \bigcup_{\rho \in T_\nu, a=1,...,M^{l-1}_\rho} MP^{l-1}_{\rho,a}$
- IF ν is an AND node
  1. Composition: $\{P^l_{\nu,b}\} = \oplus_{\rho \in T_\nu, a=1,...,M^{l-1}_\rho} MP^{l-1}_{\rho,a}$
  2. Pruning: $\{P^l_{\nu,a}\} = \{P^l_{\nu,a} | \left\langle \mathtt{w}, \Psi_\nu(P^l_{\nu,a}) \right\rangle > Thres_l\}$
  3. Local Maximum: $\{(MP^l_{\nu,a}, CL^l_{\nu,a})\} = LocalMaximum(\{P^l_{\nu,a}\}, \epsilon_W)$ where ϵ_W is the size of the window W^l_ν defined in space, orientation, and scale.

Figure 3. The inference algorithm.

als which have low fitness score ($\langle \mathtt{w}, \Psi(x,y) \rangle$ evaluated for the configuration) and use clustering which selects a small set of *max-proposals* (each representing a cluster).

The pseudo-code for the algorithm is shown in figure 3. The input to a level l is a set of max-proposals $\{MP^{l-1}_{\nu,a}\}$ for each node ν at level $l-1$ (each max-proposal, or proposal, is a configuration $\{Z^{l-1}_{\nu,a}\}$ of the subtree with root node ν and is indexed by a or b). The max-proposals generate proposals $\{P^l_{\mu,b}\}$ for nodes at level l by composition, if level l consists of AND nodes, or by union if l contains OR nodes. We prune out this set of proposals by rejecting those with low fitness scores (i.e. $\langle \mathtt{w}, \Psi(x,y) \rangle$ evaluated for the configuration) and by clustering using *local maximum* to group the proposals into a set of clusters $\{CL^l_{\mu,a}\}$, each represented by a max-proposal $\{MP^l_{\mu,a}\}$ (the local maximum is taken with respect to spatial position, scale, and orientation). The output $\{MP^l_{\mu,a}\}$ is used as input to the next level $l+1$. See [3] for full details.

5. Max Margin AND/OR Graph Learning

5.1. Primal and Dual Problems

The task of AND/OR graph learning is to estimate the parameters $\mathtt{w}$ from a set of training samples $(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ drawn from some fixed, but unknown probability distribution. In this paper, x is image, y is the configurations of an AND/OR graph.

We formulate this learning task in terms of the max-margin criterion which is designed to learn the parameters which are best for classification (i.e. to estimate y) rather than use the standard maximum likelihood criterion (see [23] for a justification for this strategy). But observe that the classification is over the set of values $\mathcal{Y}$, which is exponentially large, and hence differs greatly from simple binary classification. Effectively max-margin learning seeks to find values of the parameters $\mathtt{w}$ which ensure that the energies $\langle \Psi(x,y), \mathtt{w} \rangle$ are smallest for the ground-truth states y and for states close to the ground-truth. A practical advantages of max-margin learning is that it gives a computationally tractable learning algorithm (which avoids the need to compute the partition function of the distribution).

The main idea of the max margin approach is to forego the probabilistic interpretation of equation 1. Instead we concentrate on the discriminative function $F(x,y,\mathtt{w}) = \langle \Psi(x,y), \mathtt{w} \rangle$. We define the *margin* γ of the parameter $\mathtt{w}$ on example i as the difference between the true parse y_i and the best parse y^*:

$$\gamma_i = F(x_i, y_i, \mathtt{w}) - \max_{y \neq y_i} F(x_i, y, \mathtt{w}) \quad (2)$$

$$= \langle \mathtt{w}, \Psi_{i,y_i} - \Psi_{i,y^*} \rangle \quad (3)$$

where $\Psi_{i,y_i} = \Psi(x_i, y_i)$ and $\Psi_{i,y} = \Psi(x_i, y)$.

Intuitively, the size of margin quantifies the confidence in rejecting the incorrect parse y using the function $F(x,y,\mathtt{w})$. Larger margins [23] leads to better generalization and prevents over-fitting.

The *goal* of max margin AND/OR graph learning is to maximize the minimum margin:

$$\max_\gamma \gamma \quad (4)$$

$$s.t. \langle \mathtt{w}, \Psi_{i,y_i} - \Psi_{i,y} \rangle \geq \gamma L_{i,y}, \forall y; \|\mathtt{w}\|^2 \leq 1; \quad (5)$$

where $L_{i,y} = L(y_i, y)$ is a loss function (note there are an exponential number $|\mathcal{Y}|$ of constraints in equation 5). The purpose of the loss function is to give partial credit to states which differ from the groundtruth by only small amounts (i.e. it will encourage the energy to be small for states near the groundtruth).

The loss function is defined as follows:

$$L(y_i, y) = \sum_{\nu \in V^{AND}} \triangle(z^i_\nu, z_\nu) + \sum_{\nu \in V^{LEAF}} \triangle(z^i_\nu, z_\nu) \quad (6)$$

where $\triangle(z^i_\nu, z_\nu) = 1$ if $dist(z^i_\nu, z_\nu) \geq \delta$. Otherwise, $\triangle(z^i_\nu, z_\nu) = 0$. $dist(.,.)$ is a measure of the distance between two points and δ is a threshold. Note that the summations are defined over the active nodes. This loss function which measures the distance/cost between two parse trees is calculated by summing over individual parts. This ensures that the computational complexity of the loss function is linear in the size of the LEAF and AND nodes of the hierarchy.

By standard manipulation, the optimization can be reformulated as minimizing the constrained quadratic cost function of the weights:

$$\min_{\mathtt{w}} \frac{1}{2}\|\mathtt{w}\|^2 + C \sum_i \xi_i \quad (7)$$

$$s.t. \langle \mathtt{w}, \Psi_{i,y_i} - \Psi_{i,y} \rangle \geq L_{i,y} - \xi_i, \forall y; \quad (8)$$

where C is a fixed penalty parameter which balances the trade-off between margin size and outliers. Outliers are

training samples which are only corrected classified after using a slack variable ξ_i to "move them" to the correct side of the margin. The constraints are imposed by introducing Lagrange parameters $\alpha_{i,y}$ (one α for each constraint).

The form of the solution to this minimization can be expressed in form:

$$\mathtt{w}^* = C \sum_{i,y} \alpha^*_{i,y} \left(\Psi_{i,y_i} - \Psi_{i,y} \right), \tag{9}$$

where the α^* are obtained by maximizing the dual function:

$$\max_{\alpha} \sum_{i,y} \alpha_{i,y} L_{i,y} -$$
$$\frac{1}{2} C \sum_{i,j} \sum_{y,z} \alpha_{i,y} \alpha_{j,z} \left\langle \Psi_{i,y_i} - \Psi_{i,y}, \Psi_{j,y_j} - \Psi_{j,z} \right\rangle \tag{10}$$
$$s.t. \sum_{y} \alpha_{i,y} = 1, \forall i; \alpha_{i,y} \geq 0, \forall i, y; \tag{11}$$

Observe that the solution will only depend on the training samples (x_i, y_i) for which $\alpha_{i,y_i} \neq 0$. These are the so-called *support vectors*. They correspond to training samples that either lie directly on the margin or are outliers (that need to use slack variables). The concept of support vectors is important for the optimization algorithm that we will use to estimate the α^* (see next subsection).

It follows from equations (9,11), that the solution only depends on the data by means of the inner product $\Psi \cdot \Psi'$ of the potentials. This enables us to use the kernel trick [5] which replaces the inner product by a kernel $K(,.,)$ (interpreted as using features in higher dimensional spaces). In this paper, the kernels $K(,)$ take two forms, the linear kernel, $K(\Psi, \Psi') = \Psi \cdot \Psi'$ for image features Ψ^D and the radial basis function (RBF) kernel, $K(\Psi, \Psi') = \exp\left(-r\|\Psi - \Psi'\|^2\right)$ for shape features Ψ^H where r is a parameter of RBF.

5.2. Optimization of the Dual

The main problem with optimizing the dual, see equation 10, is the exponential number of constraints (and hence the exponential number of $\{\alpha_{i,x}\}$ to solve for). We risk having to enumerate all the parse trees $y \in \mathcal{Y}$ which is almost impractical for an AND/OR graph. Fortunately, in practice only a small number of support vectors will be needed (equivalently, only a small number of the $\{\alpha_{i,y}\}$ will be non-zero). This motivates the working set algorithm [1, 22] to optimize the objective function in equation 10. The algorithm aims at finding a small set of *active constraints* that ensure a sufficiently accurate solution. More precisely, it sequentially creates a nested working set of successively tighter relaxations using a cutting plane method. They [1, 22] show that the remaining (exponentially many) constraints are guaranteed to be violated by no more than ϵ,

Loop over k

- $y^* = \arg\max_y H(x_k, y)$ where $H(x_k, y) = \langle \mathtt{w}, \Psi_{i,y} \rangle + L(y_k, y)$.
- if $H(x_k, y^*; \alpha) - \max_{y \in S_k} H(x_k, y; \alpha) > \varepsilon$
  $S_k \leftarrow S_k \bigcup y^*$
  $\alpha_s \leftarrow$optimize dual over S, $S = S \bigcup S_k$

Figure 4. Working Set Optimization

Given a training set S and parameter α
Repeat

- select a pair of data points (y_j, y_k) not satisfying KKT conditions.
- solve optimization problem on (y_j, y_k)

Until all pairs satisfy KKT conditions.

Figure 5. Sequential Minimal Optimization

without needing to explicitly add them to the optimization problem. The pseudocode of the algorithm is given in figure 4. Note that the inference algorithm is performed at the first step of each loop. Therefore, the efficiency of training algorithm highly depends on the computational complexity of an inference algorithm (recall that we show in section 4 that the complexity of the inference algorithm is polynomial in the size of the AND/OR graph). Thus, the efficiency of inference makes the learning practical. The second step is to create the working set sequentially and then estimate the parameter α on the working set. The optimization over the working set is performed by Sequential Minimal Optimization (SMO) [12]. This involves incrementally satisfying the Karush-Kuhn-Tucker (KKT) conditions which are used to enforce the constraints. The pseudo-code is depicted in figure 5. The details of SMO is beyond the scope of this paper. See [12] for the detailed implementation.

6. Experiments

6.1. Dataset and Implementation Details

We performed the experimental evaluations using 48 human baseball images in Mori's dataset [9] as the testing set. This testing set has several advantages: (i) the ground truth of segmentation of human body and positions of key points are provided , and (ii) many results [18, 10, 9] are reported for comparisons. Some examples of the dataset are shown in figure 7 (The parsing and segmentations results are obtained by our method). Observe that the dataset contains a large variance of poses of human body and the appearance of clothes changes a lot from image to image. We created a training dataset by collecting 156 images from the internet and got students to manually label the parse tree for each image.

The AND/OR graph learnt by max-margin was used to

obtain the parse y (i.e. to locate the body parts). We used max-margin on the training dataset to learn the parameters of the max-margin model. During learning, we set $C = 0.1$ in equation 10, used the radial basis function kernel with $r = 0.1$, set the parameter in the loss function (equation 6) to be $\delta = 12$, and set $\epsilon = 0.01$ in figure 4. Our strategy to obtain segmentation, which is inspired by Grab-Cut [16], is to obtain the parse by the inference algorithm on the AND/OR graph and then segment object by graph-cut using the feature statistics inside the boundary as initializations (note that, unlike us, Grab-Cut requires initialization by a human).

6.2. Performance Comparisons

Two evaluation criterions are used to measure the performances of parsing and segmentation. The *average position error* [18] is used as the measure of the quality of parsing. The position error means the distance at pixel level between the positions of groundtruth and the parsing result. The smaller the position error, the better the quality of the parsing. Srinivasan and Shi [18] only used 5 joint nodes (head-torso, torso-left thigh, torso-right thigh, left thigh-left lower leg, right thigh-right lower leg) per image. In our case, there are 27 nodes along the boundary of human body per image used to give more detailed parsing. We use the segmentation measure, '*overlap score*' named by [18], to quantify the performance of segmentation. The overlap score is defined by $\frac{area(P \cap G)}{area(P \cup G)}$, where P is the area which the algorithm outputs as the segmentation and G is the area of ground-truth. The bigger the overlap score, the better the segmentation.

We compare the performances obtained by our approach to those reported by Srinivasan and Shi [18], which are the best results achieved so far on this dataset (e.g. better than Mori et al. 's [9]). Firstly, we compare the average position errors in figures 6. Observe that our best parse gives performance slightly better than the best (manually selected) of the top 10 parses output by [18] and significantly better than the best (manually selected) of their top three parses. Secondly, we compare the average overlap scores in figure 6. The difference of performance measured by overlap score is more significant. Observe that our result is significantly better than the best (manually selected) of their top 10 parses.

We illustrate our parsing and segmentation results in figure 7. The dotted points indicate the positions of the leaf nodes of parse tree which lie along the boundary of human body. The same parts in different images share the same color. For example, yellow and red points correspond to the left and right shoulder respectively. Light blue and dark blue points correspond to the left and right legs respectively. One can observe that the variation of poses are extremely large. Our AND/OR graph is capable of covering the articulated poses of body parts and segmenting the body nicely. However, the inference takes 3 minutes for image with size 640×480.

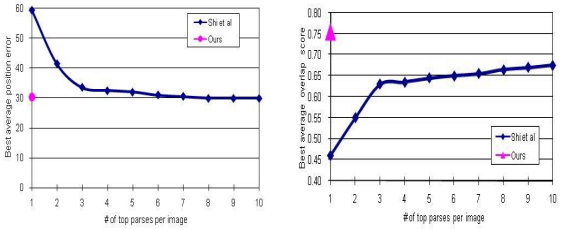

Figure 6. We compare our results with that of Srinivasan and Shi [18]. The performance of parsing (position error) and segmentation (overlap score) are shown in the top and bottom figures respectively. Note that [18] select the best one (manually)of the top parses.

7. Discussion

We presented an AND/OR graph for representing objects whose parameters can be learnt in a globally optimal way by extending max-margin learning technique developed in machine learning. Advantages of our approach include (i) the ability to model the enormous number of poses that occur for articulated objects such as humans, (ii) the discriminative power provided by max-margin learning (by contrast to MLE), and (iii) the use of the kernel trick to make use of high-dimensional features. We gave detailed experiments on the baseball datasets, showing significant improvements over the state-of-the-art method. In particular, our method outputs a single parse and not a list of possible parses. We are currently working on improving the inference speed of our algorithm by using a cascade strategy. We are also extending the model to represent humans in more detail.

8. Acknowledgments

This research was supported by NSF grant 0413214.

References

[1] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden markov support vector machines. In *ICML*, pages 3–10, 2003.

[2] H. Chen, Z. Xu, Z. Liu, and S. C. Zhu. Composite templates for cloth modeling and sketching. In *CVPR (1)*, pages 943–950, 2006.

[3] Y. Chen, L. Zhu, C. Lin, A. L. Yuille, and H. Zhang. Rapid inference on a novel and/or graph for object detection, segmentation and parsing. In *NIPS*, 2007.

[4] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[5] N. Cristianini and J. Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press, New York, NY, USA, 2000.

[6] Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. In *CVPR (2)*, pages 2145–2152, 2006.

[7] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

Figure 7. The first column shows the parse results of human body. Color points indicate the positions of body parts. The same color points in different images correspond to the same parts. The second column show the segmentations of human body. The next four columns show extra examples.

[8] M. W. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *CVPR (2)*, pages 334–341, 2004.

[9] G. Mori. Guiding model search using segmentation. In *ICCV*, pages 1417–1423, 2005.

[10] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR (2)*, pages 326–333, 2004.

[11] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *CVPR*, pages 130–136, 1997.

[12] J. C. Platt. Using analytic qp and sparseness to speed training of support vector machines. In *NIPS*, pages 557–563, 1998.

[13] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, pages 1129–1136, 2006.

[14] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV*, pages 824–831, 2005.

[15] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *ECCV (4)*, pages 700–714, 2002.

[16] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[17] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR (2)*, pages 2041–2048, 2006.

[18] P. Srinivasan and J. Shi. Bottom-up recognition and parsing of the human body. In *CVPR*, 2007.

[19] P. Srinivasan and J. Shi. Bottom-up recognition and parsing of the human body. In *EMMCVPR*, pages 153–168, 2007.

[20] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *NIPS*, 2003.

[21] B. Taskar, D. Klein, M. Collins, D. Koller, and C. Manning. Max-margin parsing. In *EMNLP*, 2004.

[22] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004.

[23] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[24] J. Zhang, J. Luo, R. T. Collins, and Y. Liu. Body localization in still images using hierarchical models and hybrid search. In *CVPR (2)*, pages 1536–1543, 2006.

[25] L. Zhu, Y. Chen, and A. L. Yuille. Unsupervised learning of a probabilistic grammar for object detection and parsing. In *NIPS*, pages 1617–1624, 2006.