Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Modeling Inter- and Intra-Part Deformations for Object Structure Parsing

Ling Cai[1] **Rongrong Ji**[1] **Wei Liu**[2] **Gang Hua**[3]
[1] Xiamen University, China
[2] IBM T. J. Watson Research Center, USA
[3] Stevens Institute of Technology, USA
{cailing.cs,jirongrong}@gmail.com weiliu@us.ibm.com ‘ ghua@stevens.edu

Abstract

Part deformation has been a longstanding challenge for object parsing, of which the primary difficulty lies in modeling the highly diverse object structures. To this end, we propose a novel structure parsing model to capture deformable object structures. The proposed model consists of two deformable layers: the top layer is an undirected graph that incorporates inter-part deformations to infer object structures; the base layer is consisted of various independent nodes to characterize local intra-part deformations. To learn this two-layer model, we design a layer-wise learning algorithm, which employs matching pursuit and belief propagation for a low computational complexity inference. Specifically, active basis sparse coding is leveraged to build the nodes at the base layer, while the edge weights are estimated by a structural support vector machine. Experimental results on two benchmark datasets (*i.e.*, faces and horses) demonstrate that the proposed model yields superior parsing performance over state-of-the-art models.

1 Introduction

Given a target object in the image, structure parsing refers to inferring the structural information, which is useful for applications such as object matching and alignment. It typically works under a coarse-to-fine manner, where bounding boxes are first detected by running a specific object detector to provide coarse information of object locations, and then fine parsing models, *e.g.*, [Fidler *et al.*, 2009], [Zhu *et al.*, 2010] and [Yang and Ramanan, 2011], are built to attain accurate object structures, parts, subparts and key points. In general, parsing of rigid objects works well by using simple geometry representations [Viola and Jones, 2001]. However, difficulties raise when facing deformable objects, like face [Felzenszwalb and Huttenlocher, 2005], pedestrian [Dalal and Triggs, 2005] and horse [Zhu *et al.*, 2010]. The key challenge lies in the difficulty to achieve a robust and unified geometry representation for deformable objects. In such a circumstance, a fine-gained analysis of object parts and detailed structures is highly required.

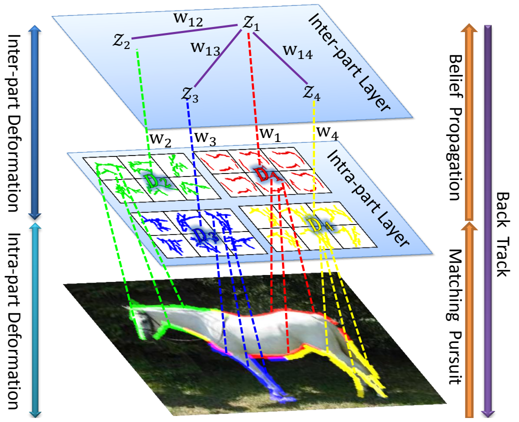


Figure 1: Two-layer deformation model for parsing object structure. The inter-part layer (the top layer) handles deformations among object parts, and the intra-part layer (the base layer) makes some local adjustments to characterize small local deformations.

In this paper, we investigate the feasibility of parsing highly deformable objects, which remains as an open problem. Specifically, we take a systematic understanding of the object deformation, and then conduct a flexible modeling for both inter- and intra-part deformations. Taking eyes and mouth in a face for instance, relative position changes between eyes and mouth belong to inter-part deformations, while appearance changes, like close eyes and open mouth, belong to intra-part deformations. Notably, most existing parsing models [Zhu *et al.*, 2010; Yang and Ramanan, 2011] are based on the independence assumption of two kinds of deformations. Typically, they concentrate on either constructing the deformable model to deal with inter-part deformations or encoding feature to comprise intra-part deformations. For modeling intra-part deformations, there exist a few approaches [Wu *et al.*, 2010], which yet cannot extract the invariance hidden in large deformations.

In this paper, we propose a novel two-layer model to jointly consider two kinds of deformations. To the best of our knowledge, there is not existing work targeting at modeling both inter-part and intra-part deformations, which is therefore unable to explore complicated object structures. As shown in

Fig. 1, it contains two layers: the base layer accounts for local intra-part deformations, which is consisted of a set of local nodes that match feature bases in small image regions to generate low-dimensional descriptors for image patches. The top layer handles inter-part deformations, which is modeled as a Markov Random Field (MRF) whose nodes correspond to the object parts of potentially large deformations, described by the feature descriptors obtained from the base layer.

The proposed two-layer model merits in a joint inference of both abstract and detailed deformation of object parts. It enables to further localize the key feature points and sketch the finer object structures. Moreover, we design a layer-wise learning algorithm which employs belief propagation for inference, which enjoys a low computational complexity by eliminating the superfluous training samples. To learn our model, active basis sparse coding is carried out to learn the nodes as bases at the base layers, while the edge weights are estimated by deploying a structural support vector machine.

The rest of this paper is organized as follows. Sec. 2 summarizes the related work. Sec. 3 builds the architecture of the proposed parsing model, including the inter- and intra-part deformation layers. Sec. 4 describes a procedure for learning the model parameters, and Sec. 5 gives an efficient algorithm for optimal structure inference. Sec. 6 presents the experimental results, and Sec. 7 draws the conclusion.

2 Related Work

Mixture models, *e.g.*, Gaussian Mixture Model (GMM) [Dempster *et al.*, 1977], are commonly used for modeling object structures. In that sense, one object can be considered as a combination of multiple conventional structure types, and the structure estimation refers to matching the unknown structure with the most similar types. Gu and Ren [Gu and Ren, 2010] employed multiple templates to detect objects in different views. In [Park *et al.*, 2010], multiple detectors at different scales are combined into a mixture model.

Compared to mixture models, tree-based models [Awasth *et al.*, 2007] are more flexible in characterizing the correlations among mixture components. A typical tree structure contains no loop, so the message passing algorithm [Kschischang *et al.*, 2001] can be applied to identify the optimal configuration. Felzenszwalb and Huttenlocher [Felzenszwalb and Huttenlocher, 2005] proposed a Pictorial Structure (PS) tree model to recognize the face structure and estimate the face pose in an image.

A branch of work has been motivated from the PS model. For example, the Articulated Model (AM) [Ramanan and Sminchisescu, 2006] employs Conditional Random Fields (CRFs) [Kumar and Hebert, 2003; Awasth *et al.*, 2007] to handle inter-part deformations. The Deformable Part Model (DPM) [Felzenszwalb *et al.*, 2010] extends the AM model to allow for more flexible spatial correlations. Nodes of DPM can be enlarged to force a more accurate parsing. For example, the Mixtures of Parts Model (MPM) in [Yang and Ramanan, 2011] exploits 26 nodes corresponding to human joints (shoulder, elbow, hand, *etc.*) for pose estimation, and the DPM in [Zhu and Ramanan, 2012] exploits 99 nodes to locate face landmarks.

However, increasing the number of nodes results in higher dimension of the encoded features as well as more training samples needed, which typically compromises the detection accuracy due to the increase of inference uncertainty. One solution is to reduce the number of parts by aggregating connected parts into a larger one. Another solution is to construct a multiple layer model [Fidler *et al.*, 2010] and allocate nodes for each layer. In this way, the high-dimensional feature space can be decomposed into multiple low-dimensional ones. In [Zhu *et al.*, 2010], Zhu *et al.* introduced the hierarchical deformable template to group every three nodes at one layer. The later hierarchical AOT model [Si and Zhu, 2013] is configured as a deep mixing of And-Or nodes to represent the compositions of parts and structural variations.

Although our proposed model is relevant to two recent multi-layer models, Hierarchical Deformable Template (HDT) [Zhu *et al.*, 2010] and And-Or Template (AOT) [Si and Zhu, 2013], our model differs from and merits over them in terms of robustly dealing with appearance variations and shape deformations. First, the HDT model [Zhu *et al.*, 2010] imposes a restriction on the number of child nodes, *i.e.*, each node can only have three child nodes at one layer. Notably, our model discards this restriction so that each node can associate an arbitrary number of child nodes. Second, unlike the independence assumption adopted in AOT [Si and Zhu, 2013], our model encodes visual appearances with geometry deformations in a joint feature encoding function, therefore allows for capturing large non-rigid deformations.

3 The Proposed Model

Considering that object structure varies significantly with poses and view points, we anticipate the proposed parsing model to deal with such variances. Unlike the HDT [Zhu *et al.*, 2010], MPM [Yang and Ramanan, 2011] and AOT [Si and Zhu, 2013], we use two deformation layers to model both inter- and intra-part deformations.As shown in Fig. 1, the top layer consists of a few nodes to represent key object parts and describes global inter-part deformations with an undirected graph. The nodes at the base layer are divided into different types of part templates, which have connections with the corresponding parent nodes at the top layer. Unlike the part detectors in [Yang and Ramanan, 2011; Zhu and Ramanan, 2012; Gu and Ren, 2010], the part templates in our model are not fixed, but are composed of many nodes which make minor adjustments independently and locally represent the distinguished sub-part regions. Such a design enables us to accurately characterize complicated intra-part deformations.

We assume that an object structure $\mathbf{Z}$ is composed of N parts, $\mathbf{Z} = (z_1, \ldots, z_N)$. Each part z_i is a node $v_i \in V$ of a graph $G = (V, E)$, and the undirected edge $e_{ij} = (v_i, v_j) \in E$ connecting two nodes encodes their pairwise interaction. In our model, z_i is denoted as a triple set $\langle x_i, y_i, t_i \rangle$ where (x_i, y_i) is its pixel location and $t_i \in \{1, 2, \ldots, T\}$ indicates which type of part template is activated, *e.g.*, the horse head may go up and down and thus has multiple types based on the head position relative to the body. Given a feature map J of the image I (J in our experiment is set as the Gabor feature map by convoluting I with rotational Gabor filters), we define a MRF to evaluate the posterior probability of an object structure configuration $\mathbf{Z}$ under the unary and pairwise potential weighting parameter set $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_N, \ldots, \mathbf{w}_{ij}, \ldots)$ and the template set $\mathbf{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_N)$,

$$p(\mathbf{Z}|J, \mathbf{w}, \mathbf{D}) \propto \exp\big(\sum_{e_{ij} \in E} h(z_i, z_j, \mathbf{w}_{ij}) + \sum_{v_i \in V} g(z_i, J, \mathbf{w}_i, \mathbf{D}_i) \big), \quad (1)$$

where the first term $h(z_i, z_j, \mathbf{w}_{ij})$ is the pairwise potential that quantifies the spatial relationship of z_i and z_j, and is responsible for modeling the inter-part deformation, as detailed in Sec. 3.1. The second term $g(z_i, J, \mathbf{w}_i, \mathbf{D}_i)$ is the unary potential that collects the local image evidence for the part z_i in J given the template $\mathbf{D}_i$ of part i. Indeed, g can be regarded as a part template to locate a certain object part in the feature map J, as detailed in Sec. 3.2. The weighting parameters $\mathbf{w}_i$ and $\mathbf{w}_{ij}$ control the contribution of g and h.

3.1 Inter-Part Deformation Layer

The inter-part deformation derives from the relative location changes among parts. Following [Yang and Ramanan, 2011; Zhu and Ramanan, 2012; Felzenszwalb *et al.*, 2010], we assume the relative location is dependent on the type t_i. That is, the relative location between them have T^2 combinations. When a part z_i is given, the other part z_j can be estimated by the posterior probability $p(z_j|z_i, \mathbf{w}_{ij})$. We rewrite $p(z_j|z_i, \mathbf{w}_{ij})$ as $p(t_j|z_i, \mathbf{w}_{ij})p(x_j, y_j|t_j, z_i, \mathbf{w}_{ij})$, and approximate it using two exponential functions, *i.e.*,

$$p(t_j|z_i, \mathbf{w}_{ij}) \propto \exp(w_{ij}^{t_i t_j (1)}), \quad (2)$$

$$p(x_j, y_j|t_j, z_i, \mathbf{w}_{ij}) \propto \exp\big(- w_{ij}^{t_i t_j (2)} (x_i - x_j - \Delta x_{ij}^{t_i t_j})^2 - w_{ij}^{t_i t_j (3)} (y_i - y_j - \Delta y_{ij}^{t_i t_j})^2 \big), \quad (3)$$

where $\mathbf{w}_{ij} = (w_{ij}^{(111)}, \ldots, w_{ij}^{T^2 (3)})$. $\Delta x_{ij}^{t_i t_j}$ and $\Delta y_{ij}^{t_i t_j}$ are the x- and y-directional mean displacements of parts z_i and z_j respectively. For simplification, we define $vx_{ij}^{t_i t_j} = -(x_i - x_j - \Delta x_{ij}^{t_i t_j})^2$ and $vy_{ij}^{t_i t_j} = -(y_i - y_j - \Delta y_{ij}^{t_i t_j})^2$ Note that Eq. 1 is an exponential function. We then define the inter-part deformation term $p(z_j|z_i, \mathbf{w}_{ij})$ by $\exp\big(h(z_i, z_j, \mathbf{w}_{ij})\big)$, *i.e.*, $h(z_i, z_j, \mathbf{w}_{ij})$ is written as

$$h(z_i, z_j, \mathbf{w}_{ij}) = w_{ij}^{t_i t_j (1)} + w_{ij}^{t_i t_j (2)} vx_{ij}^{t_i t_j} + w_{ij}^{t_i t_j (3)} vy_{ij}^{t_i t_j}, \quad (4)$$

3.2 Intra-Part Deformation Layer

In models presented in [Yang and Ramanan, 2011; Zhu and Ramanan, 2012; Gu and Ren, 2010; Park *et al.*, 2010], the invariance to intra-part deformations relies on the feature encoding scheme by using, *e.g.*, the HOG descriptor. However, the dimension of HOG is linearly related to the part's area, which would result in a high dimensional feature vector if the node at the inter-part deformation layer covers a large object part region. In addition, HOG encodes the gradient information of the block instead of the pixel, which cannot provide accurate localization information.

In contrast, active basis sparse coding [Wu *et al.*, 2010] can produce a lower dimensional feature vector by projecting image patches onto feature bases. Unlike the invariance in the HOG encoding, Matching pursuit (MP) provides another way of invariance by searching the local optimal matching for the node of template. MP has been used in [Wu *et al.*, 2010] to construct the part template and encode image patches as the sum of active bases.

In this work, we utilize MP at the intra-part deformation layer to handle local deformations and achieve the local invariance. It can be formulated as:

$$MP(J, \phi) = \max_{q \in \phi} \big(J(q)\big), \quad (5)$$

where q means a candidate from the candidate set ϕ of MP. In practice, the more candidates in ϕ, the larger deformation is allowed, yet the accuracy might decrease. The result of MP can be considered as a local image evidence of the object part z_i. It is employed as a child node of z_i at the base layer. To prevent overfitting, we restrict each part z_i to $T \times K$ child nodes at the base layer, which form T types of part templates with K members, *i.e.*, $\mathbf{D}_i = (d_i^{11}, \ldots, d_i^{TK})$. Based on the type variable t_i of z_i, the connections of K nodes of the t_i-th type are activated. Unlike the dependent nodes at the top layer, the nodes at the base layer are independent from each other for an effective representation and inference. Thus, the unary potential g of Eq. 1 is defined as the sum of contributions of K nodes:

$$g(z_i, J, \mathbf{w}_i, \mathbf{D}_i) = \sum_{k=1}^{K} w_i^{t_i k} MP\big(J, \varphi(z_i, d_i^{t_i k})\big), \quad (6)$$

where $\mathbf{w}_i = (w_i^{(11)}, \ldots, w_i^{TK})$, and $w_i^{t_i k}$ weights the contribution of the node $d_i^{t_i k}$. The function φ takes z_i and $d_i^{t_i k}$ as parameters and produces the candidate set ϕ of MP, and it is introduced in Sec. 3.3.

3.3 Feature Map

The intra-deformation layer is constructed with the feature map J of the original image I. Here, we define J as the whitening normalization of the contentional Gabor feature map introduced in [Wu *et al.*, 2010]. It normalizes the response of Gabor filtering over a number of pixels and orientation directions to make the feature map comparable among images.

The feature of a point q_k in J is denoted as $q_k = (x_k, y_k, o_k)$, where (x_k, y_k) and o_k are its spatial location and orientation, respectively. Though small deformations cause some perturbations in q_k, MP can search for the best match in the candidate set which collects feasible locations to offset these perturbations. In Eq. 6, the candidate set ϕ is generated by the function φ with two parameters z_i and $d_i^{t_i k}$, *i.e.*,

$$\begin{aligned}\varphi(z_i, d_i^{t_i k}) = \big\{ & (\tilde{x}_k, \tilde{y}_k, \tilde{o}_k); \\ & |x_i + x_i^{t_i k} - \tilde{x}_k| < \Delta d \cos \Delta o, \\ & |y_i + y_i^{t_i k} - \tilde{y}_k| < \Delta d \sin \Delta o, \\ & |o_i^{t_i k} - \tilde{o}_k| < \Delta o \big\},\end{aligned} \tag{7}$$

where $d_i^{t_i k} = (x_i^{t_i k}, y_i^{t_i k}, o_i^{t_i k})$ is the $t_i k$-th child node of z_i, and two predefined parameters $\triangle d$ and $\triangle o$ controlling the size of the candidate set.

4 Model Learning

Given the labeled positive samples $\big\{(J_1, \mathbf{Z}_1), \ldots, (J_{M^+}, \mathbf{Z}_{M^+})\big\}$ and the unlabeled negative samples $\big\{J_{M^+ +1}, \ldots, J_M\big\}$, the model learning is to estimate $\mathbf{w}$ and $\mathbf{D}$, in which the parameter set $\mathbf{w}$ balances the contributions of the unary and pairwise potentials, while the $\mathbf{D}$ localizes the object parts in the feature map. Here, we follow the assumption in [Zhu *et al.*, 2010; Yang and Ramanan, 2011; Dalal and Triggs, 2005; Zhu and Ramanan, 2012] that inter-part and intra-part deformations are independent. Therefore $\mathbf{w}$ and $\mathbf{D}$ can be learned separately from the training samples. According to the model architecture, the base layer, *i.e.*, intra-part deformation, should be learned prior to the top layer, *i.e.*, the inter-part deformation, to generate the template set $\mathbf{D}$. Then, the learning procedure in the top layer only estimates the weighting vectors $\mathbf{w}$.

4.1 Learning Intra-Part Deformation Layer

At the intra-part deformation layer, the set $\mathbf{D}$ includes $N \times T \times K$ Gabor feature bases $\big\{d_1^{(11)}, \ldots, d_N^{TK}\big\}$ as the base layer nodes. According to Eq. 7, the location of the base layer node is determined by its parent node z_i and its offset location $(x_i^{t_i k}, y_i^{t_i k})$. Thus, we collect feature maps from the same part and the same type, and search for K distinguished Gabor feature bases to construct a type of part template of $\mathbf{D}_i$. For each type of part, MP is used to search the feature map for the maximum filtering response over all locations, which pools a probability distribution. Then, projection pursuit (PP) extracts the first K feature bases with highest probabilities to form a part template. For more details about learning $\mathbf{D}_i$, the reader is encouraged to refer to [Wu *et al.*, 2010].

Notably, we assume each training sample is assigned a label z_i composed of the part location (x_i, y_i) and the part type t_i. However, it is difficult to give unambiguous definitions of the part's location and type, *i.e.*, x_i, y_i and t_i cannot be directly labeled by hand.With these key points, active shape model (ASM) [Ginneken *et al.*, 2002] is adopted to align the root part and project key points onto a uniform coordinate. Then, K-means is used to generate clusters for each object part. The center of a cluster is considered as the part location (x_i, y_i) , and the cluster label as its type t_i.

4.2 Learning Inter-Part Deformation Layer

For simplicity, we rewrite Eq. 1 as the inner product of the weighting vector $\mathbf{w}$ and the feature function $\mathbf{F}(\mathbf{Z}, J, \mathbf{D})$:

$$p\left(\mathbf{Z} | J, \mathbf{w}, \mathbf{D}\right) \propto \exp\big(\mathbf{w}^{\mathrm{T}} \mathbf{F}(\mathbf{Z}, J, \mathbf{D})\big), \tag{8}$$

where

$$\mathbf{w} = \Big[w_1^{(11)}, \ldots, w_N^{KT}, \ldots, w_{ij}^{t_i t_j (1)}, w_{ij}^{t_i t_j (2)}, \ldots\Big]^{\mathrm{T}}$$

$$\mathbf{F}(\mathbf{Z}, J, \mathbf{D}) = \Big[u_1^{(11)}, \ldots, u_N^{TK}, \ldots, 1, vx_{ij}^{t_i t_j}, \ldots\Big]^{\mathrm{T}}$$

and $u_i^{t_i k}$ denotes $MP\big(J, \varphi(z_i, d_i^{t_i k})\big)$.

Based on the structural support vector machine [Tsochantaridis *et al.*, 2005], the weighting vector $\mathbf{w}$ can be estimated by maximizing the margin between positive samples and negative samples:

$$\min \tfrac{1}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w} + \lambda \sum_{m=1}^{M} \xi_m, \tag{9}$$

$$\begin{aligned}\text{s.t.} \quad & \mathbf{w}^{\mathrm{T}} \mathbf{F}\left(\mathbf{Z}_m, J_m, \mathbf{D}\right) \geq 1 - \xi_m, \forall m \leq \mathrm{M}^+ \\ & \mathbf{w}^{\mathrm{T}} \mathbf{F}\left(\mathbf{Z}, J_m, \mathbf{D}\right) \leq -1 + \xi_m, \forall m > \mathrm{M}^+, \forall \mathbf{Z}.\end{aligned}$$

The constraints make the learned model assign large scores and small scores to positive samples and negative samples, respectively. The training samples violating these principles result in the penalty variables $\xi_m > 0$. Moreover, the cutting-plane method [Joachims *et al.*, 2009] is employed to search for those active constraints and ignore inactive ones.

5 Inference

Given the feature map J, the two-layer deformation model defined in Eq. 1 can be applied to make a confidence score for an arbitrary structure $\mathbf{Z}$. However, for a simple object with only 5 parts and 5 mixture types in a 320×240 image, the number of possible configuration is $(320 \times 240 \times 5)^5$, hence it is impractical to adopt the brute-force search. Instead, we propose to conduct an efficient inference to determine the optimal configuration $\mathbf{Z}^*$, *i.e.*,

$$\mathbf{Z}^* = \arg\max_{\mathbf{Z}} \big(\sum_{v_i \in V} g(z_i, J, \mathbf{w}_i, \mathbf{D}_i) + \sum_{e_{ij} \in E} h(z_i, z_j, \mathbf{w}_{ij}) \big). \tag{10}$$

To this end, we consider the graph G as a simple tree structure, so that belief propagation (BP) [Kschischang *et al.*, 2001; Felzenszwalb and Huttenlocher, 2005] can be applied with a linear time complexity. BP inference includes three steps: *messages passing from leaves to the root, optimal root configuration determination,* and *message backtracking to determine the optimal configuration of other parts*. Without loss of generality, we define the root part as z_1, and each non-root part z_i has only one parent part $z_{pa(i)}$. Each z_i sends a message B_i to its parent $z_{pa(i)}$, that is

$$\begin{aligned}B_i(z_{pa(i)}) = \max_{z_i} \big(& g(z_i, J, \mathbf{w}_i, \mathbf{D}_i) + h(z_i, z_{pa(i)}, \mathbf{w}_{ipa(i)}) \\ & + \sum_{c \in \{pa(c)=i\}} B_c(z_i) \big).\end{aligned} \tag{11}$$

Starting from leaf parts, every non-root part sends a message to its parent, and the propagation procedure continues until the root part receives all messages. Then, the optimal

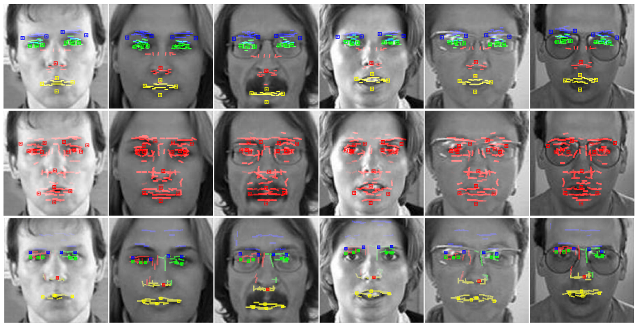


Figure 2: Parsing results on the facial images(Row 1: our model. Row 2: ABM. Row 3: AOT). The eyes, eyebrow, nose and mouth are emphasized in green, blue, red and yellow respectively. The fine color line segments represent the nodes q_i^{*k} at the base layer.

configuration z_1^* of the root part is determined as the one with the highest score, *i.e.*,

$$z_1^* = \arg\max_{z_1} \big(g(z_1, J, \mathbf{w}_1, \mathbf{D}_1) + \sum_{c\in\{pa(c)=1\}} B_c(z_1)\big), \tag{12}$$

At last, we can backtrack the messages sending to z_1^* by performing a reverse propagation, through which we can find the optimal configurations of non-root parts $z_i^* = \arg\max_{z_i}(B_i(z_{pa(i)}))$. Continuing to backtrack to the base layer, the optimal configuration of the k-th child node of part z_i^* is written as :

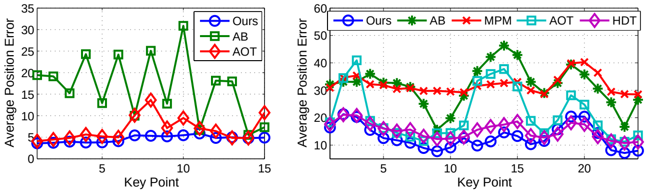
$$q_i^{*k} = \arg\max_{q\in\varphi(z_i^*, d_i^{t_ik})} \big(MP(J, \varphi(z_i^*, d_i^{t_ik}))\big), \tag{13}$$

The tree structure used in our model is simple and loop-free. We can easily extend the tree structure to a loopy graphical model. Accordingly, the loopy belief propagation [Weiss, 2000] is run to approximately infer the optimal configuration.

6 Experiments

The proposed model[1] is evaluated on Kaggle face dataset [KAG, 2013] and the Weizmann horse dataset [Borenstein and Ullman, 2008]. These two datesets provide test images with key points to quantitatively evaluate the performance of parsing models in [Zhu *et al.*, 2010; Wu *et al.*, 2010; Yang and Ramanan, 2011; Si and Zhu, 2013], especially the horse dataset which present many complicated structures, with large intra- and inter-part deformations. The effectiveness of our model is addressed through comparing to these popular and state-of-the-art deformation models. For the evaluation protocol, the average pixel distance between the predicted key points and the ground truth, namely Average Position Error (APE), is calculated for each model.

[1] MATLAB codes can be downloaded from Ling Cai's homepage at: https://sites.google.com/site/lingcai2006sjtu/parsing



(a) The Kaggle face dataset. (b) The Weizmann horse dataset.

Figure 4: The average position error of key points on two datasets.

Kaggle Face Dataset			
	Ours	AoT	AB
Eye	**3.8**	4.9	19.2
brow	**5.4**	10.1	19.7
Nose	**5.9**	7.1	6.7
Mouth	**4.9**	6.7	12.2
Total	**4.6**	6.9	16.6

(a)

Weizmann Horse Dataset					
	Ours	HDT	AoT	AB	MPM
Body	**9.9**	13.1	14.3	26.4	29.7
Head	**12.0**	15.8	29.1	38.3	31.4
FLeg	15.1	**14.5**	19.3	32.2	34.7
HLeg	**15.7**	17.4	23.7	32.3	32.3
Total	**13.0**	15.2	21.3	31.6	32.1

(b)

Table 1: The average position error of object parts on two datasets.

6.1 Face Structure Parsing

Face is composed of six basic parts (two eyes, two eyebrows, a nose and a mouth) set up in a spatial structure configuration. In the Kaggle face dataset, some parts present the intra-part deformations, *e.g.*, the mouth shape undergoes its movement when human is speaking. However, the facial expressions elicit the inter-part deformation. In addition, some accessories, like beard and eyeglasses, cause enormous difficulties for the deformation model. We test our performance on the Kaggle face dataset [KAG, 2013] to parse face images and localize the facial key points. Each image includes 15 key points of 6 facial parts. Even the same person can exhibit varying facial appearances according to his pose, the camera position or the illumination. As a result, the facial keypoint detection is actually quite challenging for the parsing models.

In our parsing model, the top layer includes 6 nodes with one type ($T = 1$) to represent the corresponding face parts, and each one has 20 child nodes at the base layer to offset the intra-part deformations. The nearest node at the base layer is used to predict the key point by adding a displacement vector. Moreover, 100 facial images are used to estimate the model parameter $\mathbf{D}$ and $\mathbf{w}$. Test is made on 250 images, and the results from our model are visualized in the first row of Fig. 2. As can be seen, for each part, the nodes at the base layer are mainly scattered around the key point to effectively encode the local region, while the node at the top layer allows the big movement to handle the inter-part deformations. Unlike two deformation layers in our model, the single intra-deformation layer in ABM cannot overcome the inter-part deformation, which generates larger positioning errors as shown in the second row. Though the grid graph in AOT can approximate the inter-part deformations, it often uses two nodes to represent one object part. Consequently intra-part deformations at the base layer are handled by the nodes at the top layer, *e.g.*, two key points at the eyebrows shown in the last row of Fig. 2.

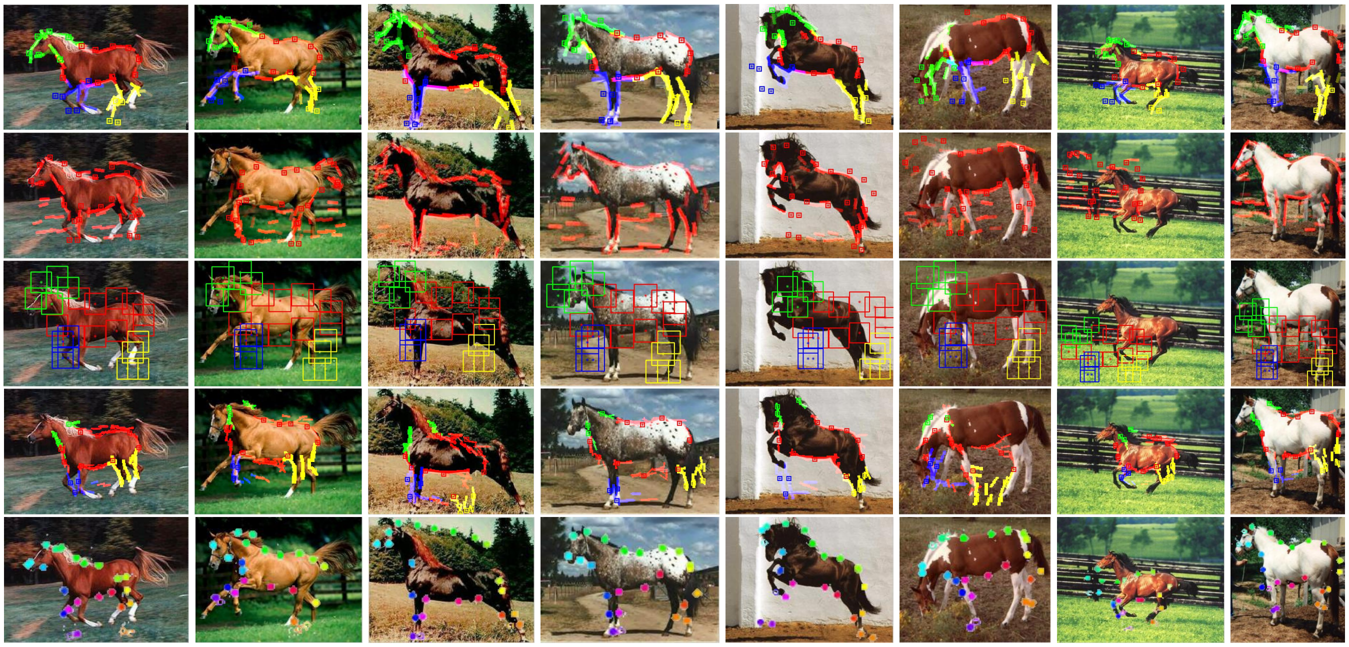


Figure 3: Horse parsing results. Row 1: our model. Row 2: ABM. Row 3: MPM. Row 4: AOT. Row 5: HDT

6.2 Horse Structure Parsing

Compared to faces, the horses in the Weizmann horse dataset [Borenstein and Ullman, 2008] present more complicated structures, with more intra- and inter-part deformations. Firstly, the structure of a horse has been represented by four parts: the root part corresponding to the horse's *body* and three child parts being the *head*, the *front legs* and the *hind legs*. Secondly, we transformed all the training images to a uniform coordinate system by aligning the root parts, so that the scales and orientations are almost the same for the training samples. At last, the clustering approach is run to determine the part label. With the clusters of the object parts using PP, we constructed a template for each type of object parts. 164 images in the dataset have been considered as positive samples to impose positive constraints for the parameter $\mathbf{w}$. To predict the specific key points on the horse, we employ the node in the intra-part layer as the anchor and a linear displacement vector to determine the positions of these points. Both ABM and our model use 80 Gabor features to represent object structure, while MPM and HDT detects 24 specific object parts as its structure.

Fig. 3 shows the comparison result of the five models for the horse parsing. The 1st row illustrates the results produced by our model, with the Gabor bases corresponding to the head, body, front legs and hind legs being drawn in green, red, blue and yellow line segments, respectively. Note that, some mixed colors appear in the overlap parts, *e.g.*, the yellow segments in the neck region is the mixture of red and green. Without the inter-part deformation, ABM (2nd row of Fig. 3) fuses the simple similarity transform to learn a single template and fits it with different object structures, so it can only represent the most common structure and the others are considered as noise, shown in the second row. HOG gains better invariance to minor local deformations than Gabor feature, but it loses the capability of accurate position detection, *e.g.*, the results in the 3rd row of Fig. 3 indicate that MPM cannot accurately determine some object parts, especially for the front and hind legs. To overcome the shortages of Gabor features, the intra-part deformation layer is constructed in our model to handle those local deformations. Similarly, the base layer in AOT deals with the intra-part deformations, but the top layer based on a grid graph present strict limitations on the part's size and position. Occasionally, an object part is represented by multiple nodes. Moreover, the grid graph does not allow the nodes for large movements. Therefore, to a certain extend, AOT is limited on describing some parts with large inter-part deformations, like head and legs as shown in the 4th row of Fig. 3. Unlike the two deformation layers in our model, HDT, based on the hierarchical structure, contains multiple deformation layers but can only include three nodes at each layers in order to construct an invariant triplet vector. As shown in the last row of Fig. 3, the structures from HDT parsing results are just composed of some key points, which is much less informative than the sparse representation produced by our model.

6.3 Quantitative Evaluation

APE is used to quantitatively evaluate the performance of different parsing models. There is no scale issue on the Kaggle face dataset [KAG, 2013] as all images have the same size, so we measure APE on the original images. Fig. 4 (a) shows the APE of 15 key points in 300 facial images. Table 1 (a) show the APE of parts and the total APE, respectively. It can be seen that our model achieves more accurate position results at almost all key points comparing against ABM and AOT, and has the total APE less than 5.

Unlike facial images, the horse images of the varying size should be noticed for the Weizmann horse dataset [Borenstein and Ullman, 2008]. To offset the scale variations, all test images are transformed to a unified coordinate system by aligning the body part. Compared to facial images, the horse images undergo larger intra- and inter-part deformations, which pose a greater challenge to the parsing model. The quantitative evaluation for the horse samples is illustrated in Fig. 4 (b). The APE of parts and the total APE of the five models are listed in Table 1 (b). It is clear that our model gives the smallest APE at 22 key points of 24 key points, and has the total APE less than 14.

7 Conclusion

In this paper, we proposed a novel two-layer object parsing model, which incorporates and jointly learns inter- and intra-part deformations towards robust object parsing. The base layer accounts for intra-part deformations, and carries out active basis sparse coding to encode an object part as a feature vector. The top layer captures inter-part deformations through employing a MRF to accommodate large deformations of object parts. By designing such a two-layer architecture, the optimal object structure can be effectively and efficiently inferred via our devised layer-wise learning algorithm. Extensive qualitative and quantitative experimental results corroborate that the proposed model can tolerate larger part deformations and characterize finer object structures than the state-of-the-art models.

Acknowledgments

This work is supported by China National Science Foundation (No. 61422210, No. 61373076 and No. 61175009). We would like to thank Dr. Fengna Wang and the anonymous reviewers for their valuable comments.

References

[Awasth *et al.*, 2007] Pranjal Awasth, Aakanksha Gagrani, and Balaraman Ravindran. Image modeling using tree structured conditional random fields. In *IJCAI*, pages 2060–2065, 2007.

[Borenstein and Ullman, 2008] Eran Borenstein and Shimon Ullman. Combined top-down/bottom-up segmentation. *IEEE Trans. PAMI*, 30(12):2109–2125, 2008.

[Dalal and Triggs, 2005] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[Dempster *et al.*, 1977] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, pages 1–38, 1977.

[Felzenszwalb and Huttenlocher, 2005] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[Felzenszwalb *et al.*, 2010] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. PAMI*, 32(9):1627–1645, 2010.

[Fidler *et al.*, 2009] Sanja Fidler, Marko Boben, and Aleš Leonardis. Learning hierarchical compositional representations of object structure. *Object Categorization: Computer and Human Vision Perspectives*, pages 196–215, 2009.

[Fidler *et al.*, 2010] Sanja Fidler, Marko Boben, and Aleš Leonardis. A coarse-to-fine taxonomy of constellations for fast multi-class object detection. In *ECCV*, pages 687–700, 2010.

[Ginneken *et al.*, 2002] Bram Van Ginneken, Alejandro F Frangi, Joes J Staal, Bart M ter Haar Romeny, and Max A Viergever. Active shape model segmentation with optimal features. *IEEE Trans. Medical Imaging*, 21(8):924–933, 2002.

[Gu and Ren, 2010] Chunhui Gu and Xiaofeng Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, pages 408–421, 2010.

[Joachims *et al.*, 2009] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[KAG, 2013] https://www.kaggle.com/c/facial-keypoints-detection, 2013.

[Kschischang *et al.*, 2001] Frank R Kschischang, Brendan J Frey, and H-A Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. IT*, 47(2):498–519, 2001.

[Kumar and Hebert, 2003] Sanjiv Kumar and Martial Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *CVPR*, pages 1150–1157, 2003.

[Park *et al.*, 2010] Dennis Park, Deva Ramanan, and Charless Fowlkes. Multiresolution models for object detection. In *ECCV*, pages 241–254, 2010.

[Ramanan and Sminchisescu, 2006] Deva Ramanan and Cristian Sminchisescu. Training deformable models for localization. In *CVPR*, pages 206–213, 2006.

[Si and Zhu, 2013] Zhangzhang Si and Song-Chun Zhu. Learning and-or templates for object modeling and recognition. *IEEE Trans. PAMI*, 35(9):2189–2205, 2013.

[Tsochantaridis *et al.*, 2005] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. In *JMLR*, pages 1453–1484, 2005.

[Viola and Jones, 2001] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages I–511, 2001.

[Weiss, 2000] Yair Weiss. Correctness of local probability propagation in graphical models with loops. *Neural computation*, 12(1):1–41, 2000.

[Wu *et al.*, 2010] Yingnian Wu, Zhangzhang Si, Haifeng Gong, and Song-Chun Zhu. Learning active basis model for object detection and recognition. *IJCV*, 90(2):198–235, 2010.

[Yang and Ramanan, 2011] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392, 2011.

[Zhu and Ramanan, 2012] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012.

[Zhu *et al.*, 2010] Long Zhu, Yuanhao Chen, and Alan Yuille. Learning a hierarchical deformable template for rapid deformable object parsing. *IEEE Trans. PAMI*, 32(6):1029–1043, 2010.