

Learning Visual Symbols for Parsing Human Poses in Images

Fang Wang^{1,2} Yi Li²

¹ Nanjing University of Science and Technology, China

² NICTA, Canberra, Australia
{fang.wang, yi.li}@nicta.com.au

Abstract

Parsing human poses in images is fundamental in extracting critical visual information for artificial intelligent agents. Our goal is to learn self-contained body part representations from images, which we call *visual symbols*, and their *symbol-wise* geometric contexts in this parsing process. Each symbol is individually learned by categorizing visual features leveraged by geometric information. In the categorization, we use Latent Support Vector Machine followed by an efficient cross validation procedure. Then, these symbols naturally define geometric contexts of body parts in a fine granularity. When the structure of the compositional parts is a tree, we derive an efficient approach to estimating human poses in images. Experiments on two large datasets suggest our approach outperforms state of the art methods.

1 Introduction

Parsing human poses in images has been a classical topic in artificial intelligence for decades [Marr, 1982]. This research facilitates a number of fundamental studies ranging from visual perception [D’Ausilio *et al.*, 2012] to computer vision [Felzenszwalb and Huttenlocher, 2005], to particularly cognitive robotics in recent years [Jenkins *et al.*, 2007].

We focus on learning visual representations of body parts in this parsing process, which we call *visual symbols*. [Marr, 1982] has already argued that any meaningful representation of the human body should be *self-contained* in a semantic hierarchy. In his work, the main ingredients of “self-containedness” are (i) self-contained unit must have a limited complexity, such that (ii) information appears in geometric contexts appropriate for recognition, and (iii) the representation can be processed flexibly.

However, recent research is contrary to this intuitive philosophy. The body parts are frequently represented by plain cardboard models (*e.g.*, joints or limbs only). Since each part is not distinctive, the visual units are considered as an approach for computing probability of the body part locations, and the geometric contexts are coarsely defined as simple distributions between parts. As such, the inference models may

have to go beyond tree structures to model long range interactions in this coarse structure (*e.g.*, [Sun *et al.*, 2012]). This essentially makes the problem less tractable, and approximate inference has to be adopted.

Can we still use exact and fast inference to remedy the problems caused by the deformable nature of human beings? We propose to use compositional parts and exploit *symbol-wise geometric contexts* map for effective pose parsing in still images. Our part representation is compositional, *i.e.*, each part may contain one or more physical joints of the human body, and the relationship between two parts can be either hierarchical (parent-child, *e.g.*, leg and upper leg, Fig. 1) or flat. This allows us to categorize distinctive visual features for body parts, which are the descriptors that characterize the properties of image patches, and eventually model pairwise interactions in fine granularity.

A popular view of the geometric context is that pairwise relationship between two parts can be characterized by a distribution. Fig. 1b shows a distribution of relative locations between upper/lower leg (denoted by green/blue colors) and “leg” (denoted by cyan), respectively.

It is seemingly legitimate to assume both point sets satisfy normal distributions, but let us take a further look at the data. Assuming we have learned the symbols of the “leg” as depicted in Fig. 1c. We can group all the instances in each point set to a few categories (Fig. 1c and 1d). Further, we redraw the relative locations of upper/lower leg that are only associated to the corresponding symbols of the leg (Fig. 1e). Clearly, we have two observations: 1) relative locations may have different distributions and exhibit different characteristics, and 2) some categories may have similar distributions, but they are compactly distributed and have much smaller variances compared to Fig. 1b. Therefore, it is much easier to model these subsets separately.

The concept of symbols, although it is not new in vision and artificial intelligence, has not been used explicitly in many state of the art pose estimation methods. Such a set of learned symbols enables encoding the symbol-wise geometric information in a finer scale, and provides more information in inference. Therefore, it is critical to learn symbols for compositional parts.

With the help of geometric information, we categorize visual features from part instances, and use cross validation to select the best categories. We used Histogram of Oriented

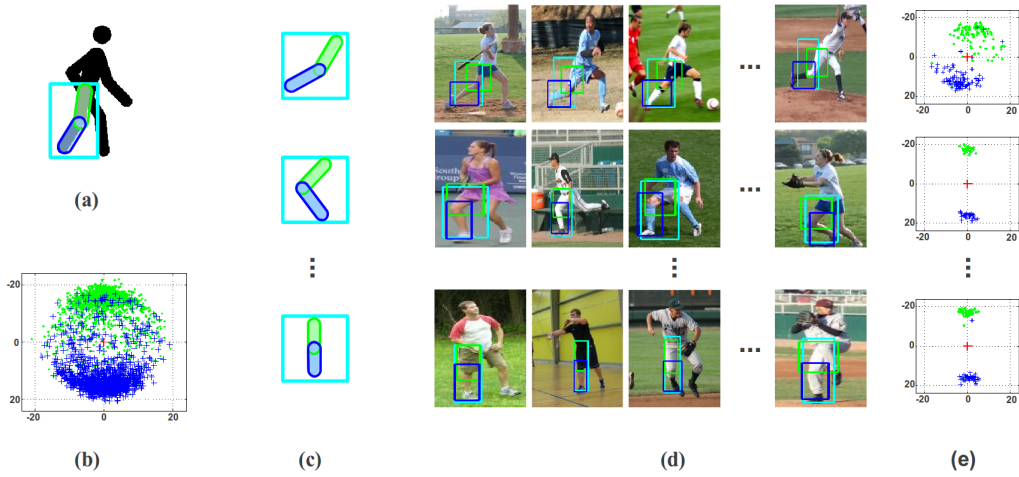


Figure 1: Our motivation. (a) human body. Green denotes upper leg, Blue denotes lower leg. (b) relative distances between upper/lower leg and leg in a large dataset, respectively. If we can group the instances to symbols (c)(d), we can easily see that relative distances can be modeled in a fine scale (e). The coordinate is defined in the image space (pixels).

Gradient (HOG) [Dalal and Triggs, 2005] in this paper. HOG is frequently used as the feature for the appearance model of body parts. In this descriptor, an image is divided into smaller regions called cells, and the histogram of gradient directions of the pixels within each cell is calculated as the descriptor.

Following human kinematic structure, we derive an approach to effectively estimate human poses. We demonstrate the performance of our method in two large datasets, and our method outperforms the state of art methods.

Our approach has the following contributions:

- We explore an effective procedure for learning self-contained symbols of body parts in parsing poses.
- Our symbol-wise context map naturally encodes the legitimate combinations of human poses from images.
- Our representation is very flexible, therefore, it is compatible with the majority of the popular inference algorithm in human pose estimation.

2 Related work

Marr is among the first to propose a hierarchical organization of body for parsing human poses [Marr, 1982]. Each model in this hierarchy is a self-contained unit, and the geometric contexts among these units are designed for recognition. His structure motivates a number of approaches in artificial intelligence, and evolves to *deformable models* in recent years, where pose estimation has been formulated as a part based inference problem.

Pictorial Structure Model (PSM) [Felzenszwalb and Huttenlocher, 2005] is one of the most successful deformable models. A tree structure graphical model is used, and pairwise terms are based on the relative distances between corresponding body parts. [Yang and Ramanan, 2011] proposed a mixtures-of-parts model for articulated pose estimation. Instead of modeling both location and orientation of body limbs as rigid parts (e.g., [Andriluka *et al.*, 2009]), they used non-oriented pictorial structures with co-occurrence constraints.

Their work relies on the geometry to define clusters (called “types” in their paper). Therefore, the representation is less *self-contained* from Marr’s point of view.

Following this research direction, [Sun and Savarese, 2011] used predefined symbols for a simultaneous detection of body parts and estimation of human poses. [Tian *et al.*, 2012] used compatibility maps in a tree structure. Latent nodes encode compatibility between parts, and accuracy is improved because incompatible poses are pruned. However, their “type” variables are solely based on the geometry, and do not encode visual information.

Beyond tree structures, graphical models were proposed in pose estimation. [Tran and Forsyth, 2010] evaluated the performance of human parsing in full relational model. A recent work by [Sun *et al.*, 2012] showed the solution of a loopy star model using Branch-and-Bound strategy. These structures usually lead to better performance, but all require efficient approximate inference methods.

All above methods used various definitions of card-board style parts, such as limbs [Andriluka *et al.*, 2009] and joints [Yang and Ramanan, 2011]. This may appear easier to model, but essentially the features are less distinctive and the performance is limited by coarse geometric context. Beyond these plain structures, researchers seek to compositional parts in recent years to characterize higher level visual representations.

[Wang *et al.*, 2011] proposed to use compositional parts to provide more precise results, but his method has a higher computation cost due to the loopy graph structure. [Rothrock and Zhu, 2011] cast human pose into AND/OR graphs, and performed human parsing using top-down scheme. Rich appearance models were adopted to estimate human parts. [Bourdev *et al.*, 2010] proposed to use poselets for human recognition. Each poselet refers to a combined part that is distinctive in training images. Please note that these poselets do not characterise geometric contexts in modeling pairwise distributions, which makes it less effective to fully capture the body dynamical structures.



Figure 2: Training visual symbols. Given a set of instances of a compositional part (a), our approach categorizes these instances (b) and summarize it to symbols by a set of linear filters (c). (d) Our tree structure model for parsing human poses. Semantically, there are two “high level” parts (red), nine “mid level” parts (orange), and 14 joints (blue) in total.

3 Approach

Define the set of M parts as $P = \{P_i\}, i \in [1, \dots, M]$. One of our goals is to learn a symbol set S such that an instance of a body part can be labelled by an entry in S .

In many state of the art datasets, the locations of primitive parts (*e.g.*, joints) are manually annotated. Therefore, the goal of our approach is to learn and assign symbols for compositional parts in the training set, and to detect compositional parts in test images by inferring their symbols.

We first present our approach to learning visual symbols for compositional parts, then we derive the compatibility map for fine scale modeling of geometric context. Finally, we adopt an efficient learning and inference method when the structure of graphical models is a tree.

3.1 Learning symbols for compositional parts

Let $G = (P, E)$ denote the relationship graph, where P denotes the body parts, and E is the set of edges that denote the pairwise relationship between parts. An instance of a compositional part $p_i = (loc_i, s_i)$, where loc_i is the image coordinate, and s_i denotes visual appearance.

We exploit the advantages of both geometric and visual information. We first use geometric information to coarsely group image patches for p_i to different clusters, then we categorize the visual features in each cluster to visual symbols. This is more computational efficient than first categorizing visual features then grouping locations to symbols.

Geometric grouping

Given a pair of connected parts (p_i, p_j) in G , we use the first part p_i as reference, and calculate the relative locations of p_j with respect to p_i . As a result, all samples of p_j are projected into 2D geometry space.

For efficiency consideration, we run k -mean to group samples in k_j geometry clusters, each of which denotes a geometry type of p_j .

Discovering visual categories

Instances for compositional parts still have large variations in appearance within a geometric type. Therefore, the geometry alone is not powerful to characterize symbols, and we need

to further learn visual categories in each geometric group for generating more discriminative visual symbols.

Fig. 2 illustrates the learning process. Given a set of instances of a part (Fig. 2a), our approach categorizes these instances to a number of subsets that are meaningful both in geometric context and visual appearance (Fig. 2b), and summarize it to symbols by a set of linear weights (Fig. 2c)

Let $\phi(I, p_i)$ denotes the visual feature of p_i in the image I . For instances of p_i within the same geometry context, we aim at learning linear classifiers to categorize visual features.

We followed [Divvala *et al.*, 2012] and built a Latent Support Vector Machine model for learning visual categories. Given N_1 positive instances of a compositional part, and N_2 negative instances, we learn K categories of this positive set. This allows us to generate the labels $L = l_1, l_2, \dots, l_{N_1+N_2}$, $l_i \in [1, K]$ for each instance. Our objective function is as follows

$$\begin{aligned} \arg \min_w \frac{1}{2} \sum_{k=1}^K \|w_k\|^2 + C \sum_{i=1}^{N_1+N_2} \epsilon_i, \\ y_i w_{l_i} \phi(p_i) \geq 1 - \epsilon_i, \epsilon_i \geq 0, \\ l_i = \arg \max_k w_k \phi(p_i) \end{aligned} \quad (1)$$

where $y_i = \{1, -1\}$ denotes whether y_i is from positive or negative sample sets, and w_k are the weights of the feature map for each part. Similar as other clustering methods, we use k -mean for initializing categorization.

Cross validation

To achieve effective training when number of visual samples is small, we turn to a cross-validation learning paradigm to discover the best classifiers and fine tune the performance. The main idea of cross validation is to perform a training step followed by a validation step [Singh *et al.*, 2012]. During the validation stage, each classifier is evaluated on the validation set, and weak classifiers with few detected samples will be removed from the classifier set.

The whole training process is conducted iteratively. Algorithm 1 illustrates the whole training process. After cross training the “survived” classifiers can be regarded as a mixture of visual symbols.

Algorithm 1 Cross validation.

Input:

H : training set for a compositional part P_i ;
 K_{in} : the number of classifiers.

Output:

K_{out} : the number of visual categories;
 w_k : linear classifiers $k = [1, \dots, K_{out}]$.

Procedure:

1. Divide the training set equally to H_1 and H_2 ;
 2. Train the classifiers $w_{[1, \dots, K_{in}]}$ on H_1 using Eq. 1;
 3. Evaluate the classification result on H_2 ;
 4. If the number of detected samples for w_i is small:
 Remove w_i from classifiers and $K_{in} = K_{in} - 1$;
 5. Swap H_1 and H_2 ;
 6. Repeat step 1-5 for t times ($t = 10$ in our experiments);
 7. $K_{out} = K_{in}$ and output $w_{[1, \dots, K_{out}]}$;
-

Discussion

The visual categorization process of a compositional part characterizes the appearance models in a way that they can be regarded as “templates”. When HOG feature is used, the set of learned weights is also considered as “HOG filters”.

In this way, our symbols encode both geometric and visual appearance information. This makes our descriptors different from other work, because they are more discriminative and representative.

3.2 Defining symbol-wise geometric context

Assigning each part a symbol allows us to build a compatibility map for any pair of symbols. Assume we have two compositional parts p_i and p_j , each of which has symbols s_i and s_j , respectively, we create the pairwise compatibility term between parts as follows.

$$D(I, p_i, p_j) = \omega_{ij}^{s_i s_j} \psi_{ij}(p_i, p_j) + b_{ij}^{s_i s_j}, \quad (2)$$

where $\psi_{ij}(p_i, p_j) = [dx, dy, dx^2, dy^2]$ denotes the relative distance between p_i and p_j , $\omega_{ij}^{s_i s_j}$ denotes the symbol-specific weights, and $b_{ij}^{s_i s_j}$ denotes the bias of the compatibility. If two symbols are not “compatible”, *i.e.*, they never exist in any training image together, this bias term is $-\infty$. Both terms are learned in Sec. 3.3.

In graphical models, we frequently model the energy minimization problem by passing messages from one node to the other. This message passing step is practical in both exact inference or inexact inference. We can utilize this compatibility term, which result in fine scale message passing. For node P_j , the incoming message $m_{k \rightarrow j}(p_j, s_j)$ from other nodes P_k and the outgoing message $m_j(p_j, s_j)$ are computed as:

$$m_j(p_j, s_j) = \sum_{k \in n(j)} m_{k \rightarrow j}(p_j, s_j) \quad (3)$$

$$m_{k \rightarrow j}(p_j, s_j) = \max_{s_k} \left[\max_{p_k} [m_k(p_k, s_k) + D(I, p_k, p_j)] \right] \quad (4)$$

where $n(j)$ denotes the neighbors of p_j in G .

3.3 Inference

Given a set of visual symbols, compatibility map, and their graph G , one can learn the parameters and perform inference. When the structure of G is a tree, the inference is exact.

In our experiment, we define the following compositional parts (Fig. 2d). Semantically, our structure has three levels: “upper body” and “lower body” as a coarse modeling of the human body, head, upper and lower limbs used in midlevel description, and joints in the fine level.

Appearance term We use HOG templates to represent each visual subcategory. For each part p_i in an image I , the appearance score of a local patch can be written as [Yang and Ramanan, 2011]

$$B(I, p_i) = \omega_i^{s_i} \phi(I, p_i), \quad (5)$$

where $\omega_i^{s_i}$ is the weight for symbol s_i in the i^{th} part. This term can be initialized by the results w_k in Eq. 1.

Deformable term Pairwise term between p_i, p_j is defined as the symbol-wise context (Eq. 2). This term can be computed effectively by distance transform in inference.

Objective function Our objective function is as follows

$$p = \arg \max_p \sum B(I, p_i) + \sum_{i,j} D(I, p_i, p_j) \quad (6)$$

Since our model is a tree, standard message passing algorithm (Eq. 3 and Eq. 4) and exact inference are applicable.

Learning model parameters The objective function (Eq. 6) can be rewritten as $f = \langle \theta, \Phi \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product, θ consists of image filters for single parts ($\omega_i^{s_i}$), pairwise deformable weights ($\omega_{ij}^{s_i s_j}$) and biases $b_{ij}^{s_i s_j}$, and $\Phi(I, p)$ denotes the concatenated features from appearance and deformable components. The learning of θ amounts to the quadratic optimization:

$$\arg \min_{\theta, \xi_i \geq 0} \frac{1}{2} \theta^T \cdot \theta + C \sum_{n=1}^N \xi_i, \quad (7)$$

$$\forall n \in \text{pos} \quad \theta^T \Phi(I_n, p) \geq 1 - \xi_n,$$

$$\forall n \in \text{neg} \quad \theta^T \Phi(I_n, p) \leq -1 + \xi_n.$$

This is a standard quadratic programming procedure, and can be solved effectively.

4 Experiments

We present our experiments in this section. First, we describe the datasets we used for evaluation. Then, we demonstrate the visual symbols learned by our method, and compare our approach against four other methods.

In our experiments, we extract HOG features on grid image with 4×4 pixels from image patches, and learn visual symbols and geometric context map. The number of geometric cluster k_j is 8 for large compositional parts, and 6 for small

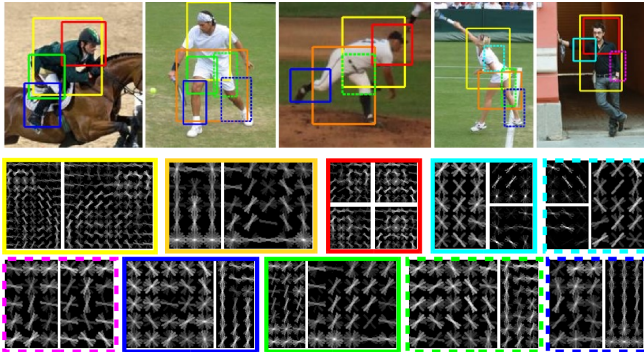


Figure 3: We show filters for our compositional parts: torso (yellow), lower body (orange), head (red), upper arm (cyan), lower arm (magenta), upper leg (green), lower leg (blue). Left/right side are denoted by solid/dashed lines, respectively.

parts, and the number of visual symbols for each geometric types is set to 2 and 4, respectively. The number of geometric clusters is consistent with that in [Yang and Ramanan, 2011]. The final number of the appearance clusters depends on the cross validation. As a result, we learn 8 to 20 visual categories for each visual symbol after cross validation.

4.1 Dataset

We evaluate our performance on two large datasets, namely, Image Parse dataset [Ramanan, 2007] and Leeds Sport dataset [Johnson and Everingham, 2010]. In all experiments, we used 500 images in the negative set of INRIA person dataset as negative samples ([Dalal and Triggs, 2005])

Image Parse dataset

Image Parse dataset (PARSE) contains 305 images with annotated poses. This dataset has images from various human activities, background and different illuminations. All images are resized such that human in images have roughly the same scale (150 pixels in height). We used 100 images for training, and the rest for testing.

LSP dataset

The recent Leeds Sport Dataset (LSP) contains 2000 images. This collection has a larger variation of pose changes. Humans in each image were cropped and scaled to 100 to 150 pixels in height. The dataset was split evenly into training set and testing set.

4.2 Demonstration

We demonstrate the effectiveness of learning procedure in this section. Fig. 3 shows localization results for the eleven higher level compositional parts, as well as examples of their filters learned by our method. In this visualization, we use different colors and line types to denote different parts.

Filters for visual symbols Each filter in Fig. 3 exhibits a few characteristics for the corresponding compositional part. These filters are related, in the sense that they model the human body at different levels. Each filter is also self-contained, *i.e.*, stand alone part detectors for different tasks.

This intrinsic constraint facilitates the inference. For instance, the torso (yellow) filters indicate the body inclination in a coarse level, which limits the search of head position (red) both in geometric context and appearance.

Interpretation The localization results can be regarded as multi level symbolic annotations for the human body. Therefore, this can be used for a number of applications ranging from high level understanding to low level description. We can further assign semantic meaning to these symbols. For instance, the torso and lower body locations can be used for analyzing the actions of the human beings (“stretching”, “standing”, “squat”, etc), and the midlevel limb detection results are used for motion analysis (“extending arm”, “fetching”, etc.).

4.3 Comparison

We compared our approach against four state of the art methods for human poses parsing on the PARSE and the LSP dataset in this section. In our comparison, we used the criterion in [Ferrari *et al.*, 2008] for performance evaluation. A part is correctly detected if both its endpoints are within 50% of the length of corresponding ground truth segments. Then we used the probability of a correct pose (PCP) to measure the percentage of correctly localized body parts.

Table 1 summarizes the evaluation results, with highest scores being highlighted. We compared the parsing accuracy of our method against [Andriluka *et al.*, 2009], [Yang and Ramanan, 2011], [Johnson and Everingham, 2010], and [Tian *et al.*, 2012], respectively. We re-run the method of [Andriluka *et al.*, 2009] and [Yang and Ramanan, 2011] on the datasets and report the results, and other data entries in the table are from the original papers, respectively. Please note that [Tian *et al.*, 2012] tried two different settings in their methods, but they did not report the result using all the 1000 images in the training set.

Overall our method achieved the best performance. In the PARSE dataset, our method is marginally better than the original algorithm (75.7% vs 74.9%). This is possibly because the power of our visual category training may not be fully explored due to small number of training samples.

The problem caused by limited training samples is relieved in the LSP dataset. When 1000 images are used for training, our method outperforms other methods. Compared to four methods, our total detection accuracy (65.2%) is consistently higher. Our performance is also superior to [Johnson and Everingham, 2011], where 11000 samples were used for training¹. This suggests that our training is effective.

Fig. 4 shows some examples of parsing results for three methods. Each triplet contains results for [Andriluka *et al.*, 2009], [Yang and Ramanan, 2011] and ours. The parsing results show that our method produces visually pleasing poses. The interaction between high level and low level compositional parts makes our results more “balanced”. For instance, detection results for two legs are well separated if the corresponding symbols of the lower body part are detected, be-

¹Due to a large number of missing labels in this dataset, we do not perform our evaluation on it. In their method, the training samples were automatically relabeled during optimization.



Figure 4: Result comparison. Each triplet of image contains results for [Andriluka *et al.*, 2009] in its original visualization, [Yang and Ramanan, 2011] and ours using the visualization scheme in [Johnson and Everingham, 2011] .

Exp.	Method	Torso	Head	Upper Leg	Lower Leg	U.Arm	L.Arm	Total
PARSE	[Yang and Ramanan, 2011]	97.6	93.2	83.9	75.1	72.0	48.3	74.9
	Ours	97.1	90.2	86.1	77.1	74.9	46.9	75.7
LSP	[Yang and Ramanan, 2011]	92.6	87.4	66.4	57.7	50.0	30.4	58.9
	[Tian <i>et al.</i> , 2012] (first 200)	93.7	86.5	68.0	57.8	49.0	29.2	58.8
	[Tian <i>et al.</i> , 2012] (5 models)	95.8	87.8	69.9	60.0	51.9	32.9	61.3
	[Johnson and Everingham, 2010]	78.1	62.9	65.8	58.8	47.4	32.9	55.1
	[Johnson and Everingham, 2011]	88.1	74.6	74.5	66.5	53.7	37.5	62.7
	[Andriluka <i>et al.</i> , 2009]	76.8	68.5	56.9	48.8	37.4	20.0	47.1
	Ours	92.2	84.7	78.1	67.5	54.7	37.2	65.2

Table 1: Performance on the PARSE and the LSP dataset. The first two rows shows the performance comparison on the PARSE dataset against [Yang and Ramanan, 2011]. The next seven rows show the performance of four algorithms on the more challenging LSP dataset.

cause the symbol-wise geometric context naturally guides the maximization (Eq. 6) to this optimal solution. Our method also tries to make the best guess of self-occluded parts.

The training takes approximately 8 hours on a 2.8GHz Quad Core CPU with 6GB memory. Test takes 1.5s for an 320×240 image. Compared to models where loop BP is used, our tree structure essentially speeds up the training. The running time of our method is in the same order of magnitude of [Yang and Ramanan, 2011]. Therefore, our method strikes a balance between accuracy from long range interaction and the efficiency from exact inference.

5 Conclusion

This paper presents a novel approach to learning self-contained representations for parsing human poses in images. The main contribution is the visual symbols that facilitate geometric context modeling. Our method can be used for many graphical models. When the model is a tree, we demonstrate that our method outperforms four current methods.

References

- [Andriluka *et al.*, 2009] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.
- [Bourdev *et al.*, 2010] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. *ECCV*, 2010.
- [Dalal and Triggs, 2005] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005.
- [D’Ausilio *et al.*, 2012] A. D’Ausilio, L. Badino, Y. Li, S. Tokay, L. Craighero, R. Canto, Y. Aloimonos, and L. Fadiga. Leadership in orchestra emerges from the causal relationships of movement kinematics. *PLoS ONE*, 7(5):e35757, 05 2012.
- [Divvala *et al.*, 2012] S. Divvala, A. Efros, and M. Hebert. How important are deformable parts in the deformable parts model? *CoRR*, abs/1206.3714, 2012.

- [Felzenszwalb and Huttenlocher, 2005] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61(1):55–79, January 2005.
- [Ferrari *et al.*, 2008] V. Ferrari, M. Marín-Jiménez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.
- [Jenkins *et al.*, 2007] O. Jenkins, G. Serrano, and M. Loper. Interactive human pose and action recognition using dynamical motion primitives. *I. J. Humanoid Robotics*, 4(2):365–385, 2007.
- [Johnson and Everingham, 2010] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. doi:10.5244/C.24.12.
- [Johnson and Everingham, 2011] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, pages 1465–1472, 2011.
- [Marr, 1982] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA, 1982.
- [Ramanan, 2007] D. Ramanan. Learning to parse images of articulated bodies. *NIPS*, 19:1129, 2007.
- [Rothrock and Zhu, 2011] B. Rothrock and S.C. Zhu. Human parsing using stochastic and-or grammars and rich appearances. In *Computer Vision Workshops (ICCV Workshops)*, pages 640–647. IEEE, 2011.
- [Singh *et al.*, 2012] S. Singh, A. Gupta, and A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, Berlin, Heidelberg, 2012. Springer-Verlag.
- [Sun and Savarese, 2011] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, 2011.
- [Sun *et al.*, 2012] M. Sun, M. Telaprolu, H. Lee, and S. Savarese. An efficient branch-and-bound algorithm for optimal human pose estimation. In *CVPR*, pages 1616–1623, 2012.
- [Tian *et al.*, 2012] Y. Tian, C. L. Zitnick, and S. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *ECCV*, 2012.
- [Tran and Forsyth, 2010] D. Tran and D. Forsyth. Improved human parsing with a full relational model. *ECCV*, 2010.
- [Wang *et al.*, 2011] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR*, 2011.
- [Yang and Ramanan, 2011] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.