A Latent Clothing Attribute Approach for Human Pose Estimation

Weipeng Zhang[†], Jie Shen[†], Guangcan Liu[‡], Yong Yu[†]

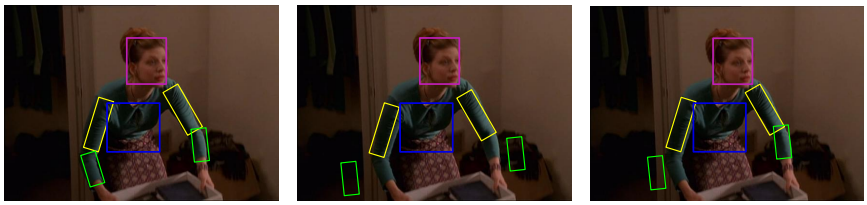
[†]Shanghai Jiao Tong University, [‡]Nanjing University of Information Science and Technology

Abstract. As a fundamental technique that concerns several vision tasks such as image parsing, action recognition and clothing retrieval, human pose estimation (HPE) has been extensively investigated in recent years. To achieve accurate and reliable estimation of the human pose, it is well-recognized that the clothing attributes are useful and should be utilized properly. Most previous approaches, however, require to manually annotate the clothing attributes and are therefore very costly. In this paper, we shall propose and explore a *latent* clothing attribute approach for HPE. Unlike previous approaches, our approach models the clothing attributes as latent variables and thus requires no explicit labeling for the clothing attributes. The inference of the latent variables are accomplished by utilizing the framework of latent structured support vector machines (LSSVM). We employ the strategy of *alternating direction* to train the LSSVM model: In each iteration, one kind of variables (e.g., human pose or clothing attribute) are fixed and the others are optimized. Our extensive experiments on two real-world benchmarks show the state-of-the-art performance of our proposed approach.

1 Introduction

Human oriented technology has a central role in computer vision and can greatly advance daily-life related applications. For example, face verification for surveillance [1] and clothing parsing for fashion search [2]. One of the most fundamental human oriented techniques is the well-known *human pose estimation* (HPE) in 2D images. In general, HPE could facilitate many applications, e.g., action recognition [3], image segmentation [4], etc. However, it is difficult to accurately estimate the human pose in unconstrained environments, especially in the presence of vision occlusions and background clutters.

To tackle the challenges, it is well-recognized that the contextual information (e.g., clothing attributes) is useful, as illustrated in Figure 1. As a consequence, the so-called *context modeling*, which is to model properly the contextual information possibly existing in images, is widely regarded as a promising direction for HPE. A variety of approaches have been proposed and investigated in the literature over several years, e.g., [5, 4, 6]. In [5], it was proposed a model that encourages high contrast between background and foreground. Ladicky et al. [4] combined together pose estimation and image segmentation, aiming to take the



(a) (b) (c)

Fig. 1: **Examples to demonstrate the benefit of integrating clothing attributes into HPE.** In the three results of HPE, all human poses in (b) and (c) are correct except lower arms. we can assume that (c) is incorrect based on the great appearance difference between left and right lower arm, but there is slight appearance difference in (b). If we know the clothing attribute type, e.g. the sleeve type or color, we can remove (b) based on the inconsistent color between the upper and lower arms. Finally, we get the correct estimation (a).

advantages of joint learning. In [6], a unified structured learning procedure was adopted to predict human pose and garment attribute simultaneously.

While effectual, the existing approaches require to label lots of contextual messages for training, and thus they are time-consuming and impractical. In this paper, we shall introduce a *latent* clothing attribute approach for HPE. Our approach formulates the HPE problem by extending the pictorial structure framework [7, 8] and, in particular, models the clothing attributes as *latent variables*. Comparing to the previous approaches that rely on label information, our latent approach, in sharp contrast, requires no explicit labels of the clothing attributes and can therefore be executed in an efficient way. We define some clothing attributes and build their connections with human parts (e.g., sleeve with arms). Some domain specific features, including *pose-specific* features and *pose-attribute* features, are designed to describe the connections. We utilize the latent structured support vector machines (LSSVM) for the training procedure, where the attribute values are initialized by a simple K-Means clustering algorithm. Then the model parameters are learnt by employing a relabel strategy, which minimizes the objective function of LSSVM in an "alternating direction" manner. More precisely, we perform an iterative scheme to train the model: Given the (latent) clothing attributes, we perform a dynamic programming algorithm to find a suboptimal solution for human pose; Given the human pose, we seek the optimal attribute values by performing a greedy search on the attribute space. We empirically show that our approach can achieve the state-of-the-art performance on two benchmarks.

In summary, the contributions of this paper are three-folds: (1) We establish a latent clothing attribute approach that can implicitly utilize clothing attributes to enhance HPE. (2) We propose some domain specific features to describe the connections between human parts and clothing attributes. (3) We introduce an

efficient algorithm to solve the optimization problem which is indeed challenging due to the presence of latent variables.

2 Related Work

As aforementioned, HPE is a difficult problem, especially in unconstrained scenes. Some of the researchers studied the problem under the context of 3D scenery [9, 10]. In the work of [9], they extended the popular 2D pictorial structure [7, 8] to 3D images and employed the new framework to model view point, joint angle, etc. Shotton et al. [11] proposed a real time algorithm for estimating the 3D human pose, striving for making the technique practical in real world applications.

Most studies (including this work) on HPE focus on 2D static images. In the early works, the human part was often modeled by oriented template. Although straightforward, the oriented templates may not properly handle the fore-shortening of the objects [12–14]. In [15], an advanced representation scheme was proposed to model the oriented human parts. The new model is formulated as a mixture of non-oriented components, each of which is attributed with a "type". Interestingly, the new model can approximate the fore-shortening by tuning the adjacent components in a spring structure.

Some work tried to incorporate "side" techniques, e.g., image segmentation, to enhance HPE. In [16], a variety of image features, e.g., boundary response and region segmentation, were utilized to produce more reliable HPE results. In [5], the background was modeled as a Gaussian distribution. In [17], the authors present a two-stage approximate scheme to improve the accuracy of estimating lower arms in videos. The algorithm was imposed to output the candidates with high contrast to the surroundings.

Besides of the shape feature which is very discriminative, the appearance feature (e.g. color, texture) is also important for HPE [18]. Generally, the appearance feature is actually a description of the clothing. As illustrated in Figure 1, there is a strong correlation between human pose and clothing attribute. Some previous work such as [19, 2, 20, 21] utilized the result of HPE to predict the clothing attribute or retrieve similar garments. Other methods (e.g., [22, 6]) attempted to refine the clothing parsing by HPE and, in turn, refine HPE by clothing parsing. However, this requires a large annotation for clothing. In our work, it is not required to manually annotate the attributes as we take them as latent variables.

There is some work that has investigated clothing attributes in the tasks other than HPE. In [23], Liu et al. aimed to recommend garment for specific scenes. To bridge the gap between the low-level image evidence and the garment recommendation, they integrated an attribute-level representation that propagates semantic messages to the recommendation system. In [3], similar attribute techniques as ours were used for action recognition. However, there is a key difference: In [3], the attribute is used as a middle level prior and the high level task was facilitated by the knowledge of attribute; In our work, the attribute

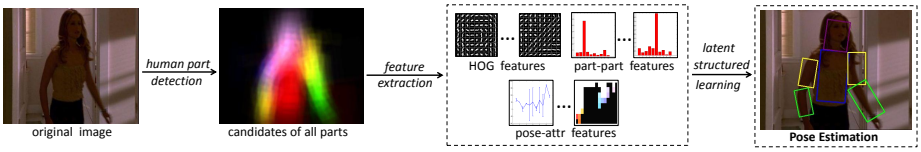


Fig. 2: **Overview of our approach.**

is modeled in a unified manner with human pose. Our model takes a relabel strategy to alternatively optimize the variables of the attribute and pose.

3 HPE with Latent Clothing Attributes

We summarize the pipeline of our approach in Figure 2. First, we take a pre-processing step to detect potential human parts in the image. This step allows us to have a search space with manageable size. Then, we extract the domain specific features to characterize the human pose and clothing attributes. Finally, we utilize the LSSVM to actualize our attribute aware human pose model and present an efficient inference algorithm to find an approximate optimal solution to LSSVM. Note that our model can reveal the clothing attributes, and thus humans with similar attribute values will be grouped together (i.e., clustering human by their clothing attributes).

Table 1: The configuration of clothing attributes

Attribute	Human parts	Features	Number of values
Sleeve	All arms	Color Histogram	3
Neckline	Torso + Head	HOG	4
Pattern	Torso	LBP [24]	5

Before introducing the proposed approach in detail, we would like to introduce some notations. We write I for an image. A human part is represented as a bounding box (x, y, s, θ), where (x, y) is the coordinate, s is the size and θ is the rotation. To obtain an input space with manageable size, we use the existing HPE method [15] to produce 40 candidates for each human part. Thus, the input space $\mathcal{X}$ of our approach is defined as:

$$\mathcal{X} = \{\mathbf{x} | \mathbf{x} = (\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m)\}, \tag{1}$$

where m is the number of human upper-body parts ($m = 6$ in this work), and $\mathbf{b}_i$ denotes the candidate ensemble for the i-th human part (there are 40 candidates in each $\mathbf{b}_i$). The output space of human pose is defined as as follows:

$$\mathcal{P} = \{\mathbf{p} | \mathbf{p} = (p_1, p_2, \cdots, p_m), \forall i, 1 \leq p_i \leq 40\}, \tag{2}$$

where p_i is a positive integer that indicates the index of the estimated candidate.

We aim to integrate clothing attributes into HPE task, striving for capturing the strong correlation between human parts and clothing attributes. We consider three types of attributes in this work, including "Neckline", "Pattern" and "Sleeve". Each attribute has multiple styles, e.g., short sleeve and long sleeve for the "Sleeve" attribute. Heuristically, for each r-th attribute ($r = 1, 2, 3$), the number of attribute values, T_r, are determined as in Table 1 (see the last column). Then the output space of the latent clothing attributes is as follows:

$$\mathcal{A} = \{\mathbf{a} | \mathbf{a} = (a_1, a_2, \cdots, a_n), \forall r, 1 \leq a_r \leq T_r\}. \tag{3}$$

where n is the number of clothing attributes ($n = 3$ in this work), and a_r is the label for the r-th attribute. Note here that it has no specific consideration to choose the value for a_r, e.g., $a_1 = 1$ may mean short sleeve or long sleeve. In this work it is an unsupervised clustering procedure that recognizes the clothing attributes.

Finally, the task of jointly estimating clothing attribute and human pose is formulated as follows:

$$f : \mathcal{X} \rightarrow \mathcal{Y}, \tag{4}$$

where $\mathcal{Y}$ is the output space given by

$$\mathcal{Y} = \{\mathbf{y} | \mathbf{y} = (\mathbf{p}, \mathbf{a}), \mathbf{p} \in \mathcal{P}, \mathbf{a} \in \mathcal{A}\}. \tag{5}$$

Regarding the prediction function f, we presume that there is a score function S which measures the fitness between any input-output pair $(\mathbf{x}, \mathbf{y})$ such that:

$$S(\mathbf{x}, \mathbf{y}; \beta) = \langle \beta, J(\mathbf{x}, \mathbf{y}) \rangle \tag{6}$$

where $\langle \cdot \rangle$ denotes the inner product between two vectors, $J(\cdot, \cdot)$ is the feature representation, and β is an unknown weight vector. In this way, the mapping function f in Eq. 4 can be written as:

$$f(\mathbf{x}; \beta) = \arg\max_{\mathbf{y} \in \mathcal{Y}} S(\mathbf{x}, \mathbf{y}; \beta) \tag{7}$$

This is a latent structured learning problem, where the latent variables are clothing attributes. Our learning procedure is motivated by [25], which employs a relabel strategy to increasingly improve the prediction of latent variables. Yet before proceeding to the training pipeline, we firstly introduce the design of the domain-specific features, as shown in the next section.

3.1 Feature Representation

The joint feature representation is an important component in structured learning [26]. We define the joint feature function $J(\mathbf{x}, \mathbf{y})$ by using two types of features, including *pose-specific* features denoted by $j_p(\mathbf{x}, \mathbf{p})$, and *pose-attribute* features denoted by $j_{pa}(\mathbf{x}, \mathbf{y})$; that is,

$$\langle \beta, J(\mathbf{x}, \mathbf{y}) \rangle = \langle \beta_p, j_p(\mathbf{x}, \mathbf{p}) \rangle + \langle \beta_{pa}, j_{pa}(\mathbf{x}, \mathbf{y}) \rangle \tag{8}$$

In the following, we present our techniques used to design each type of feature.

Pose-specific Features Given an input sample $\mathbf{x}$, we use the Histogram of Oriented Gradients (HOG) [27] to describe the shape of a candidate and consider the deformation constraint between two connected parts:

$$j_p(\mathbf{x}, \mathbf{p}) = \sum_{i=1}^{m} hog(\mathbf{x}, p_i) + \sum_{(i,j)\in E_p} d(\mathbf{x}, p_i, p_j), \tag{9}$$

where E_p is the set of connected limbs. The design of the deformation feature $d(\mathbf{x}, p_i, p_j)$ involves some basic geometry constraints between connected parts, including relative position, rotation and distance of part candidate p_i with respect to p_j, which is computed as $[x_j - x_i, y_j - y_i, (x_j - x_i)^2, (y_j - y_i)^2]$ [15].

Pose-Attribute Features Now we try to integrate the clothing attributes into our model. Notice that an attribute is only associated with some of the human parts). For a given attribute r, we denote the human parts associated with it as r_p and the corresponding configuration as P_r. The detailed inter-dependency between human parts and clothing attributes is shown in the second column of Table 1. According to the work [2], for different attributes, different low-level features should be used to achieve good performance. The specific features used for each clothing attribute can be found in the third column in Table 1.

Formally, the pose-attribute features are defined as:

$$j_{pa}(\mathbf{x}, \mathbf{y}) = \sum_{r=1}^{n} \Psi(\mathbf{x}, P_r, a_r) \tag{10}$$

where $\Psi(\mathbf{x}, P_r, a_r)$ denotes the features extracted from the human part $\mathbf{x}$, with the configuration P_r and the attribute label a_r.

Algorithm 1 Structured Learning with Latent SVM

Input: Positive samples, negative samples, initial model β, number of relabel iteration t_1, number of hard negative mining iteration t_2.
Output: Final Model β^*.

```
Initialize the final model: β* = β.
Let the negative sample set F_n = ∅.
for relabel = 1 to t_1 do
    Let the positive sample set F_p = ∅.
    Add positive samples to F_p.
    for iter = 1 to t_2 do
        Add negative samples to F_n.
        β* := Pegasos(β*, F_p ⋃ F_n).
        Remove easy negative samples:
        Remove the samples whose feature vector v satisfying ⟨β*, v⟩ < −1 from F_n.
    end for
end for
```

Similar to [6], the pose-attribute feature is designed by an outer product of low-level features and an identity vector. We first convert the clothing attribute label a_r to a T_r-dimensional vector, denoted as $L(a_r)$, one element of which is assigned with valued "1" and all others are set to be "0". From Table 1, the low-level feature descriptors of the r-th clothing attribute depend on two aspects: 1) the corresponding human parts and 2) the feature type (denoted by F_r and has been specified in Table 1). We use $F_r(P_r)$ to denote features of the r-th clothing attribute associated with the part configuration P_r. Then our pose-attribute feature $\Psi(\mathbf{x}, P_r, a_r)$ is designed as follows:

$$\Psi_{pa}(\mathbf{x}, P_r, a_r) = F_r(P_r) \otimes L(a_r) \tag{11}$$

where the "$\otimes$" operator represents the (vectorized) outer product of two vectors.

3.2 Structured Learning with Latent SVM

Now we consider the problem of learning the prediction mapping f, given a collection of images labeled with human part locations. This is the type of data available in the all standard benchmark dataset for human pose estimation. Note that clothing attributes have no labels, and we treat them as latent variables.

We describe a framework for initializing the structure of a joint model and learning all parameters. Parameter learning is done by constructing a LSSVM training problem. We train the LSSVM using the relabel approach (details will be described later) together with the data-mining (hard negative mining), and we use Pegasos [28] for the online update to solve the problem of huge space for negative samples.

Algorithm 2 Inference for Clothing Attributes

Input: A sample $\mathbf{x}$, Model parameter β , Human parts label $\mathbf{p}$
Output: optimal clothing attributes value $\mathbf{a}^*$
1: let T_r is the number of r-th clothing attribute type
2: **for** r:= 1 **to** 3 **do**
3: select the attribute value which has highest score:
$\mathbf{a}_r = \arg\max_{1 \le r \le T_r} \langle \beta^r_{pa}, j_{pa}(\mathbf{x}, P_r, a_r) \rangle$
4: **end for**

Objective Function We aim to learn the fitness function $S(\mathbf{x}, \mathbf{y}; \beta)$ defined in Eq. (6), which can later be used for joint estimation (see Eq. (7)). Given a positive training sample $(\mathbf{x}, \mathbf{y})$, we expect $S(\mathbf{x}, \mathbf{y}; \beta) \ge 1$. On the other hand, if a training sample $(\mathbf{x}, \mathbf{y})$ is negative, the output of the fitness function is required to be less than -1. In this way, given a training set $D = \{(\mathbf{x}_1, \mathbf{y}_1, z_1), \cdots, (\mathbf{x}_q, \mathbf{y}_q, z_q)\}$, where $z_k \in \{1, -1\}$ indicates the k-th sample is positive or not, we can optimize

the following objective function to solve β:

$$\min_{\beta} \frac{1}{2}\|\beta\|^2 + C\sum_{k=1}^{q} \max(0, 1 - z_k S(\mathbf{x}_k, \mathbf{y}_k; \beta)). \tag{12}$$

Initialization Since the clothing attributes are latent variables, we can only access the label of human pose. To start up, we take a relabel strategy to update the positive samples (more accurately, the clothing attribute labels) and the weight vector β in an alternative manner.

There are many ways to initialize the latent variables. One can randomly assign labels for training samples which may be unstable. In our work, we first use the groundtruth of human pose to extract low-level features (see Table 1) for each attribute. Then we perform a K-Means clustering algorithm to obtain the center of each attribute value, where K is exactly the number of attribute values we defined in Table 1. In this way, the initial label for the clothing attribute can be determined by the closest center.

Now all of the labels have been generated, we can solve Problem (12) to obtain the initial weight vector β (line 1 in Algorithm 1).

Relabel Strategy As the initial clothing attribute labels are not accurate, we employ a relabel strategy to update the attribute labels. That is, given the model parameter β and human pose, we predict the clothing attribute by maximizing the fitness function $S(\mathbf{x}, \mathbf{y}; \beta)$, which is shown in Algorithm 2. Note that according to the design of our joint feature $J(\mathbf{x}, \mathbf{y})$, the pose-specific features are irrelevant for the inference of attributes. From Eq. (10), we know that there is no interaction between different attributes since j_{pa} is summation of n separate attributes associated features. Therefore, we can perform an efficient greedy search for each attribute to obtain a local optima (line 2–4 in Algorithm 2).

Algorithm 3 Approximate Inference for Clothing Attribute Aware HPE Task

Input: A sample $\mathbf{x}$, Model parameter β.
Output: Optimal estimation $\mathbf{y}^*$ and score $S*$.
1: Set $\mathbf{y}^* = \emptyset$.
2: Set the optimal score $S^* = -\infty$.
3: Initialize the parts estimation $\mathbf{p}_0$.
4: **repeat**
5: Compute the local optimal clothing attributes $\mathbf{a}_t$.
6: Compute the local optimal human pose $\mathbf{p}_t$.
7: Compute the local score: $S = S(\mathbf{x}, \mathbf{y}_t; \beta)$.
8: **if** $S > S^*$ **then**
9: $S^* = S$, $\mathbf{y}^* = \mathbf{y}_t$
10: **end if**
11: **until** S^* not change

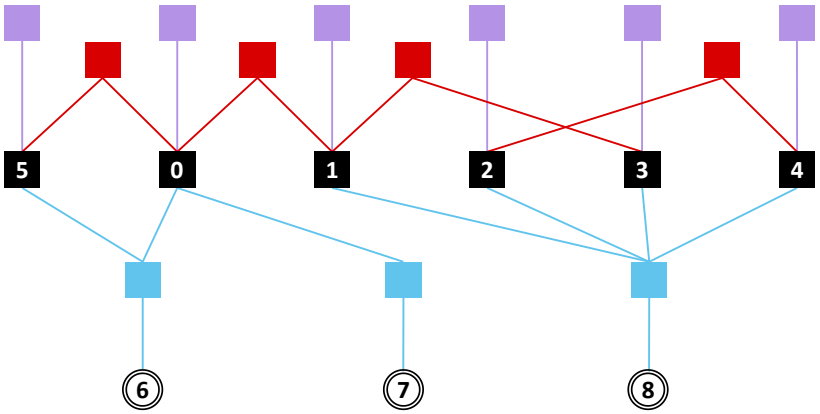


Fig. 3: **Nodes with numbers from 0 to 5 are the human part variable and those 6 to 8 are clothing attributes. Colored nodes are the potentials.**

Hard Negative Mining For a recognition or detection task, one can obtain a positive sample set with manageable size. However, there is a huge space for the negative samples. Actually, it is not possible for enumerate *all* negative samples. Thus, it is important to feed an algorithm with "hard" negative samples for efficiency and memory cost. In line 6–10 of Algorithm 1, we perform hard negative mining [25] to obtain valuable negative samples. This schema will call the inference algorithm 3 (see Section 3.3). More concretely, given an input sample $\mathbf{x}$ and weight vector β, we launch Algorithm 3 to find the optimal estimation $\mathbf{y}^*$. If $z \cdot S^*$ is less than -1 (a threshold we set), $\mathbf{x}$ is considered hard. The searching procedure on $\mathbf{x}$ will be stopped only when the S^* is greater than -1 (the $\mathbf{y}^*$ produced by the previous step is removed from the search space).

After collecting all the hard negative samples, we update β with Pegasos solver [28] (line 8 in Algorithm 1). Then we use the updated β to perform a shrinkage step to remove the easy negatives from the hard negative set F_n.

3.3 Inference

In Figure 3, we represent our problem as a factor graph $\mathcal{G}$, where the rectangle node denotes a human part, the circle node with double boundaries denotes a clothing attribute. As our original problem is a cyclic graph, it cannot be optimized exactly and efficiently. Therefore, in Algorithm 3, we propose an iterative algorithm to search for an approximate solution. Our algorithm receives a sample x, the model parameter β as inputs and outputs a local optima for human parts and clothing attribute. In each iteration, by fixing the attributes, the inference can be performed on a tree structure, which can be optimized with a dynamic programming [8]. When the human parts are fixed, an efficient greedy search schema for clothing attribute is employed (see Algorithm 2).

Algorithm 4 Inference for Human Pose

Input: A sample $\mathbf{x}$, Model parameter β , Clothing attributes value $\mathbf{a}$
Output: optimal human parts estimation $\mathbf{p}^*$

1: set the optimal human parts estimation $\mathbf{p}^* = \emptyset$
2: set the node 0 as the root node
3: **for** each candidate $\mathbf{p}_i$ of node i **do**
4: set $m(\mathbf{p}_i) = \langle \beta_p^i, \phi_p(\mathbf{x}, p_i) \rangle + \langle \beta_{pa}^r, \Psi_{pa}(\mathbf{x}, P_r, a_r) \rangle$
5: **end for**
6: **for** each candidate $\mathbf{p}_j$ of parent node j and $\mathbf{p}_i$ of child node i **do**
7: set $l(\mathbf{p}_i, \mathbf{p}_j) = \langle \beta_p^{ij}, \psi_p(\mathbf{x}, p_i, p_j) \rangle$
8: **if** i is a leaf node **then**
9: $B_i(\mathbf{p}_j) = \max_{\mathbf{p}_i}(m(\mathbf{p}_i) + l(\mathbf{p}_i, \mathbf{p}_j))$
10: **else**
11: $B_i(\mathbf{p}_j) = \max_{\mathbf{p}_i}(m(\mathbf{p}_i) + l(\mathbf{p}_i, \mathbf{p}_j) + \sum_{v \in C_i} B_v(\mathbf{p}_i))$
12: **end if**
13: **end for**
14: select the best candidate for the root node:
$\mathbf{p}_0^* = \arg\max_{\mathbf{p}_0}(m(\mathbf{p}_0) + \sum_{v \in C_0} B_v(\mathbf{p}_0))$
15: **for** each parent-child pair $(\mathbf{p}_j^*, \mathbf{p}_i)$ **do**
16: $\mathbf{p}_i^* = \arg\max_{\mathbf{p}_i} B_i(\mathbf{p}_j^*)$
17: **end for**

Inference for Human Pose We elaborate the inference procedure of human pose by extending the pictorial structure framework. In Figure 3, we denote our score with colored nodes, with purple and red ones denoting the appearance and deformation scores. The main extension for the traditional PS model is the cyan nodes, which denoting the score to measure the fitness of human pose and clothing attribute (called pose-attribute score). Therefore, we propose the human pose inference procedure in Algorithm 4. We denote the children nodes as C_i for a node i. We compute the appearance and pose-attribute scores in line 3–5. In line 7, we compute the deformation score for each parent-child pair node i and j. In the line 8–12, we compute conventional message passing procedure by dynamic programming [7]. Then we perform a top-down process to find the best candidate for each human part in line 14–17.

4 Experiments

4.1 Datasets

We evaluate our approach using the Buffy dataset [29] and the DL (daily life) dataset. The Buffy Dataset contains 748 pose-annotated video frames from Buffy TV show. This dataset is presented as a benchmark for HPE task. The DL dataset contains 997 daily life photos collected from the Flickr website. We annotate the human pose for this dataset. Compared with Buffy, the DL dataset has more various clothing attribute values. In order to obtain quantitative evaluation results for attributes, we manually annotate the clothing attributes for

Buffy and DL. There is a standard partition of Buffy for training and testing, where the training set consists of 472 images and the remaining are used for testing. For the DL dataset, we select randomly 297 images for training and use the remaining 700 images for testing.

Table 2: Comparison with State-of-the-art Algorithms on the Buffy Dataset

Method	Torso	Upper arms	Lower arms	Head	Total
Andriluka et al. [30]	90.7	79.3	41.2	95.5	73.5
Sapp et al. [16]	**100**	95.3	63.0	96.2	85.5
Yang and Ramanan [15]	**100**	96.6	70.9	**99.6**	89.1
Our Approach	**100**	**97.1**	**78.4**	99.1	**91.6**

Table 3: Comparison with State-of-the-art Algorithms on the DL Dataset

Method	Torso	Upper arms	Lower arms	Head	Total
Andriluka et al. [30]	97.0	91.7	84.5	94.0	90.6
Sapp et al. [16]	**100**	88.5	78.0	87.6	86.8
Yang and Ramanan [15]	99.8	95.7	87.5	95.6	93.6
Our Approach	**100**	**97.2**	**91.3**	**99.1**	**95.7**

4.2 Baselines and Metric

We compare our approach with three state-of-the-art algorithms: Andriluka et al. [30], Sapp et al. [16], Yang and Ramanan [15]. For the HPE results, we evaluate them with a standardized evaluation protocol based on the probability of correct pose (PCP) [31], which measures the percentage of correctly localized human parts. For the clothing attributes results, we evaluate them with a standardized metric (F1 score) of clustering task. We use the K-Means clustering results as our baseline for clothing attributes. First we use the groundtruth of human pose to obtain the clustering center for each attribute value. Then we perform K-Means clustering under a given pose, which is produced by either the state-of-the-art HPE algorithms or the groundtruth.

4.3 Results

Figure 6 shows some exemplar HPE results produced by our approach. We provide the PCP evaluation results on Buffy and DL in Table 2 and Table 3 respectively. For the Buffy dataset, Table 2 shows that our approach consistently outperforms Yang and Ramanan [15] which is a recently established algorithm.

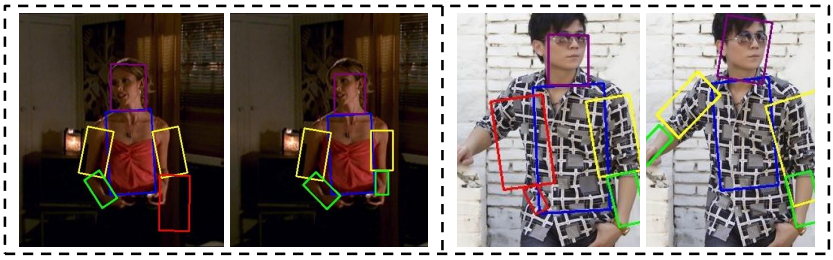


Fig. 4: **Comparison of our approach with Yang and Ramanan [15]** Yang and Ramanan [15] produces incorrect estimation (the 1st and 3rd) for upper and lower arms, while our latent clothing attribute approach produces correct.

It is expected that the most difficult parts to estimate are the lower arms. Surprisingly, the improvement on the lower arms of our approach achieves 7.5 percent higher than Yang and Ramanan, possibly because of the integration of the sleeve attribute. For the DL dataset, our algorithm consistently outperforms all the competing baselines since the photos in DL are collected from daily life and have richer clothing attributes than Buffy.

Table 4: F1 scores for clothing attributes results on Buffy

HPE	Sleeve	Neckline	Pattern	Total
Andriluka et al. [30] + K-Means	24.1	26.6	34.2	28.3
Sapp et al. [16] + K-Means	22.9	27.9	40.5	30.4
Yang and Ramanan [15] + K-Means	38.3	25.7	22.6	28.9
Groundtruth + K-Means	34.7	36.1	39.5	36.8
Our Approach	**55.6**	**68.8**	**80.8**	**68.4**

Table 5: F1 scores for clothing attributes results on DL

HPE	Sleeve	Neckline	Pattern	Total
Andriluka et al. [30] + K-Means	27.5	31.7	27.6	28.9
Sapp et al. [16] + K-Means	34.9	30.5	23.8	29.7
Yang and Ramanan [15] + K-Means	43.2	28.6	35.8	35.9
Groundtruth + K-Means	31	29.8	26.1	28.9
Our Approach	**57.2**	**60.3**	**74.7**	**64.1**

As we also aim to reveal the clothing attribute, we show some results in Figure 5 for Buffy and DL, where we arrange the images with same attribute value into one group (i.e. clustering humans by their clothing attributes). In the top pane of Figure 5, we group humans by the sleeve attribute. The performance

Fig. 5: **Examples grouped on sleeve from Buffy and neckline from DL.** The first row of the top panel (sleeve) shows the sleeveless type, the second is long type, while the first row of the bottom panel (neckline) shows the pointed type, the second is round type. The right two columns are the incorrect results.

under the F1 score is demonstrated in Table 4 and 5. Surprisingly, our approach enjoys a significant improvement on both datasets, mainly because of the relabel strategy and the iterative update role for our model parameter. Note that the result of "K-means + Groundtruth" provides the initial labels for the clothing attributes. In this way, we examine the effectiveness of our relabel strategy.

5 Conclusion

Inspired by the strong correlation between human pose and clothing attributes, we propose a latent clothing attribute approach for HPE, incorporating the clothing attributes into the traditional HPE model as latent variables. Compared with previous work [6], our formulation is more suitable for practical applications as we do not need to annotate the clothing attributes. We utilize the LSSVM to learn all the parameters by employing a relabel strategy. To start up, we take a simple K-Means step to initialize the latent variables and then update the model and the clothing attributes in an alternative manner. Finally, we propose an approximate inference schema to iteratively find an increasingly better solution. The experimental results justify the effectiveness of our relabel strategy and show the state-of-the-art performance for HPE.

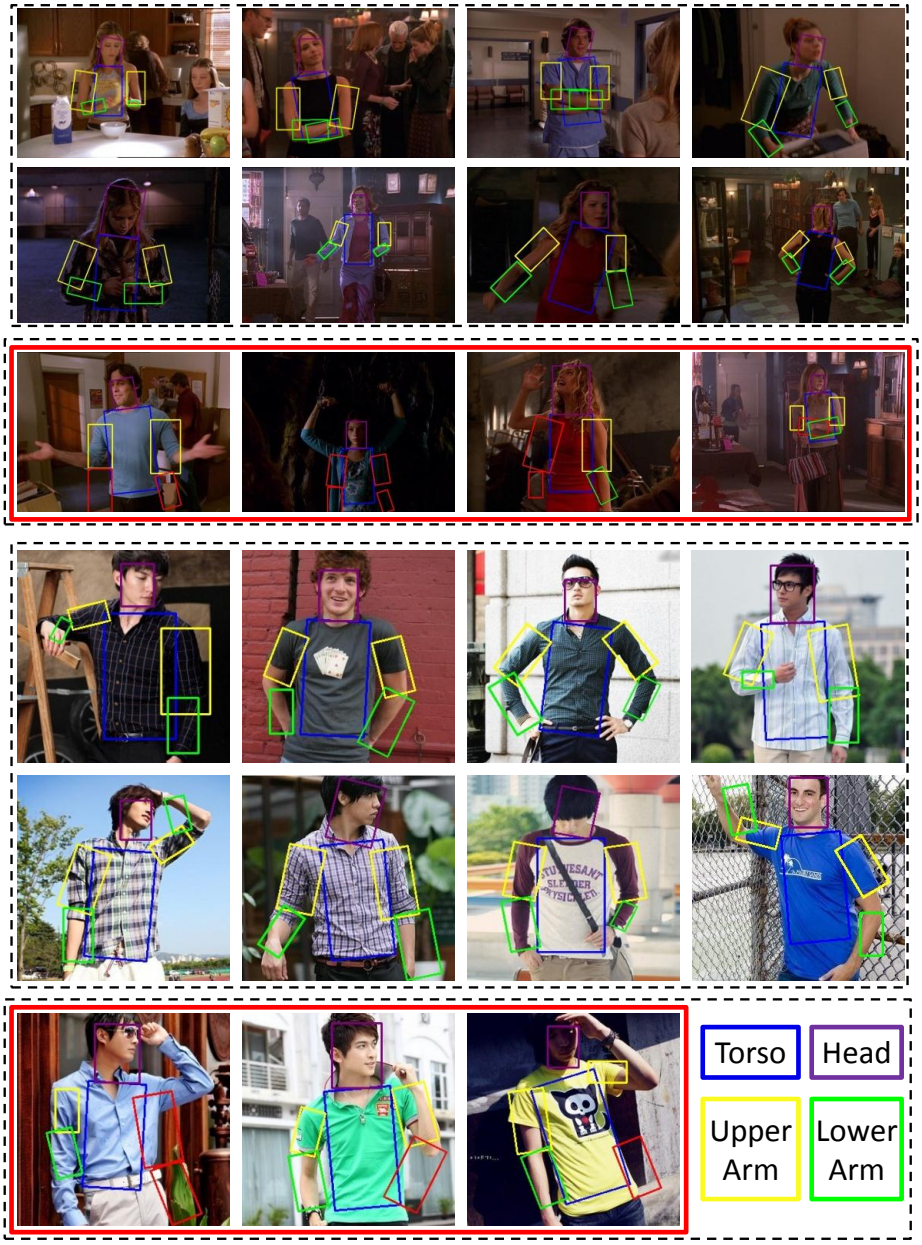


Fig. 6: **Visualization of pose results produced by our algorithm on the Buffy and DL datasets.** The top two panels are from Buffy and the others are from DL. We use the oriented bounding box to denote the pose estimation. The first panel of each dataset are correct results, while the second panel are incorrect results. The bounding box with red color denote the incorrect estimation.

References

1. Liu, L., Zhang, L., Liu, H., Yan, S.: Towards large-population face identification in unconstrained videos. In: IEEE Transactions on Circuits and Systems for Video Technology. (2014) 1
2. Liu, S., Song, Z., Liu, G., Xu, C., Lu, H., Yan, S.: Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In: IEEE Conference on Computer Vision and Pattern Recognition. (2012) 3330–3337
3. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: IEEE Conference on Computer Vision and Pattern Recognition. (2011) 3337–3344
4. Ladicky, L., Torr, P.H.S., Zisserman, A.: Human pose estimation using a joint pixel-wise and part-wise formulation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2013) 3578–3585
5. Rothrock, B., Park, S., Zhu, S.C.: Integrating grammar and segmentation for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2013) 3214–3221
6. Shen, J., Liu, G., Chen, J., Fang, Y., Xie, J., Yu, Y., Yan, S.: Unified structured learning for simultaneous human pose estimation and garment attribute classification. arXiv preprint arXiv:1404.4923 (2014)
7. Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. IEEE Transactions on Computers (1973) 67–92
8. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. International Journal of Computer Vision (2005) 55–79
9. Burenius, M., Sullivan, J., Carlsson, S.: 3d pictorial structures for multiple view articulated pose estimation. In: IEEE Conference on Computer Vision Pattern Recognition. (2013) 3618–3625
10. Ionescu, C., Carreira, J., Sminchisescu, C.: Iterated second-order label sensitive pooling for 3d human pose estimation. In: IEEE Conference on Computer Vision Pattern Recognition. (2014)
11. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Communications of the ACM (2013) 116–124
12. Ramanan, D.: Learning to parse images of articulated bodies. In: Neural Information Processing Systems. (2006) 1129–1136
13. Sapp, B., Jordan, C., Taskar, B.: Adaptive pose priors for pictorial structures. In: IEEE Conference on Computer Vision and Pattern Recognition. (2010) 422–429
14. Morris, D.D., Rehg, J.M.: Singularity analysis for articulated object tracking. In: Conference on Computer Vision and Pattern Recognition (CVPR). (1998) 289–296
15. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixture-of-parts. In: IEEE Conference on Computer Vision and Pattern Recognition. (2011) 1385–1392
16. Sapp, B., Toshev, A., Taskar, B.: Cascaded models for articulated pose estimation. In: European Conference on Computer Vision. (2010) 406–420
17. Cherian, A., Mairal, J., Alahari, K., Schmid, C.: Mixing body-part sequences for human pose estimation. In: IEEE Conference on Computer Vision Pattern Recognition. (2014)
18. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. In: British Machine Vision Conference. (2009)
19. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: European Conference on Computer Vision. (2012) 609–623

20. Bourdev, L., Maji, S., Malik, J.: Describing people: Poselet-based attribute classification. In: International Conference on Computer Vision (ICCV). (2011)
21. Li, Y., Zhou, Y., Yan, J., Niu, Z., Yang, J.: Visual saliency based on conditional entropy. In: ACCV 2009. (2009) 246–257
22. Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L.: You are what you wear: Parsing clothing in fashion photos. In: IEEE Conference on Computer Vision and Pattern Recognition. (2012) 3570–3577
23. Liu, S., Feng, J., Song, Z., Zhang, T., Lu, H., Xu, C., Yan, S.: Hi, magic closet, tell me what to wear! In: ACM Multimedia Conference. (2012) 619–628
24. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: International Conference on Pattern Recognition. (1994)
25. Felzenszwalb, P., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2010) 1627–1645
26. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research (2005) 1453–1484
27. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition. (2005)
28. Shalev-Shwartz, S., Singer, Y., Srebro, N.: Pegasos: Primal estimated sub-gradient solver for svm. In: International Conference on Machine Learning. (2007) 807–814
29. Ferrari, V., Marin, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008)
30. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2009) 1014–1021
31. Ferrari, V., Marn-Jimnez, M.J., Zisserman, A.: Pose search: Retrieving people using their pose. In: IEEE Conference on Computer Vision and Pattern Recognition. (2009) 1–8