# Parsing fashion image into mid-level semantic parts based on chain-CRFs

**Article** *in* IET Image Processing · February 2016

DOI: 10.1049/iet-ipr.2015.0507



**4 authors**, including:

Fan Wang
Beihang University (BUAA)
**4** PUBLICATIONS **5** CITATIONS
SEE PROFILE

Qiyang Zhao
Beihang University (BUAA)
**22** PUBLICATIONS **46** CITATIONS
SEE PROFILE

IET Image Processing

*Research Article*

IET Journals
The Institution of Engineering and Technology

# Parsing fashion image into mid-level semantic parts based on chain-conditional random fields

ISSN 1751-9659
Received on 1st February 2015
Revised on 29th November 2015
Accepted on 30th January 2016
doi: 10.1049/iet-ipr.2015.0507
www.ietdl.org

*Wang Fan*[1] ✉, *Zhao Qiyang*[1], *Yin Baolin*[1], *Xu Tao*[2]

[1]*State Key Laboratory of Software Development Environment, Beihang University, Beijing, People's Republic of China*
[2]*Shandong Provincial Key Laboratory of Network based Intelligent Computing, University of Jinan, Jinan, People's Republic of China*
✉ *E-mail: wangfan@nlsde.buaa.edu.cn*

**Abstract:** In this study, the authors address the problem of parsing fashion images into mid-level semantic parts including upper-clothing, lower-clothing, skin, hair and background. These mid-level parts provide the regional information of fashion items and have potential value in high-level parsing process. The key idea of the method is to parse the mid-level parts by region expanding. Owing to the co-occurrence of pose skeleton and the proposed parts, the region expanding process starts from the super-pixels crossed by specific segments of pose skeleton. The super-pixels are then merged with their neighbours by conditional inference based on their position and perceptual similarity. To avoid the difficulties of training on arbitrary graph structures, conditional random fields (CRFs) are constructed on super-pixel chains, which are extracted from the generated expanding trees. This is followed by a voting stage to mix up the probabilities estimated by the chain-CRFs to obtain the final result. Experiments on two datasets show that the new method outperforms related approaches in regional accuracy and has good generalisation capability. Furthermore, the method can be easily employed to improve the performance of high-level parsing. Its effectiveness has been verified by another group of experiments on two state-of-the-art high-level parsing approaches.

## 1 Introduction

Fashion image parsing aims at labelling images into non-overlapping semantic parts with a predefined label set. It is one of the fundamental tasks for understanding the clothes in photographs. Accompanied by other tasks including clothing segmentation [1–4] and garment attribute classification [5–7] these researches are widely used in visual applications such as clothing retrieval [8, 9], clothing recommendation [10, 11] and human recognition [12–15].

Much previous work on fashion parsing [16–18] labels pixels or super-pixels with high-level semantic concepts directly such as *dress*, *blouse* and *jeans*. However, a huge gap exists between low-level super-pixels and high-level semantic concepts. Super-pixels are usually too small to be of sufficient discriminative meanings for recognising high-level concepts, which are also mentioned in [19]. Therefore, as shown in Fig. 1, the mid-level parts are introduced in this paper. These mid-level parts indicate the regions of clothing items and provide more appearance information for recognition. Furthermore, their weak semantic meanings can be applied to reduce the number of candidate category labels. Taking Fig. 1 as an example, super-pixels from 'upper-clothing' part can be labelled as 'coat' or 'shirt', but not 'pants' or 'skirt'. In [10], clothing attributes are used as mid-level features for modelling people's preferences for clothes. Proposing a new method for clothing attribute classification and applying the attributes to improving the performance of clothing recommendation are their main contributions. Our paper also works on mid-level study, but the mid-level parts introduced refer to the regions of independent clothing items with weak semantic meanings. We focus on their regional accuracy and applying them to improve the performance of fashion parsing. To our best knowledge, no such work has been proposed in fashion parsing. Unlike clothing items of different categories having discriminative characteristics in their overall appearances, the mid-level parts distinguish each other mainly by their position and self-inner perceptual similarity. Such information can be caught by local features which are the key sources in our mid-level work.

Fashion parsing is relevant to conventional image parsing tasks such as scene parsing and object segmentation [20–22] and sharing the same major question: how to take context information into account to improve the performance. The answer to this question differs according to the type of context information. In fashion parsing, due to the co-occurrence of body parts and clothing items, human pose [23–25] is widely used. Unlike the global spatial prior and bounding box used in the other tasks [4, 26, 27], human pose is much more structured and containing rich semantic information. Making good use of human pose in fashion parsing remains a challenging problem and lots of attempts have been made such as generating effective pose features [16, 17] and reducing the impact of incorrect pose [18]. Different from these approaches, our method not only applies human pose for positioning, but further extracts structure information by generating coarse masks of clothes.

Conditional random fields (CRFs) are powerful models to deal with conditional inferences, which are widely used in fashion parsing [16, 18] and lots of visual tasks [28–31]. One representative work from Yamaguchi *et al.* [16] locate super-pixels by calculating their normalised two-dimensional coordinates to the joints of human poses and labels the super-pixels based on CRFs. Jammalamadaka *et al.* [18] expand their parsing work to unrestricted images, where probabilistic pose estimation is used to avoid parsing with a single wrong pose. Both of the studies construct CRFs model based on a graph structure with all adjacent super-pixels connected. However, fashion images vary dramatically in clothing appearance, style and background, which lead to large diversity in size and connectivity of the underlying graph structure. It is difficult to regularise the model parameters in such arbitrary structure. This is also mentioned in [32]. Unlike the two approaches, our method performs inferences on super-pixel chains generated in the region expanding process. Owing to the simple and regularised structure, training the model is much easier.

Another work from Yamaguchi *et al.* [17] solves the labelling problem by a retrieval-based approach. Their key idea is to label query images by looking at clothes with similar appearance in a

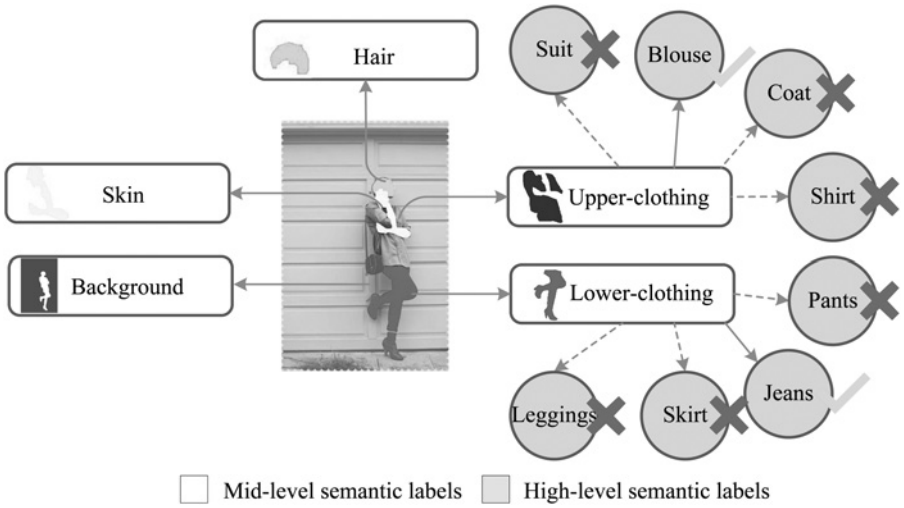


**Fig. 1** *Fashion image can be divided into five mid-level parts according to the content including hair, upper-clothing, lower-clothing, skin and background. These mid-level parts bear high-level semantic meanings. For example, the upper-clothing is indeed a blouse in the above example image*

huge retrieval dataset (330 K pictures). This method heavily relies on the retrieval precision. Owing to the large variation in the appearance of clothes, the method will not work well on clothes whose pattern or style seldom appears in the retrieval dataset. The retrieval process is also very time consuming. It takes 20–40 s to parse a single image. In contrast, our method deals with the variation by selecting multiple seed regions to represent a complicated area. These seeds are then expanded based on an efficient chain-CRFs model, which is much more light-weighted.

Several deep learning approaches have been proposed on human parsing and clothes recognition, which are related to our work. In human parsing, Luo *et al.* [33] propose a deep de-compositional network to solve the occlusion problem. Liang *et al.* [34] formulate the parsing process as active template regression. Liu *et al.* [35] map body parts from nearest neighbours to query image with matching convolutional neural network. The major issue in these researches is segmenting body parts. The handled clothing types are very limited. Our work focuses on parsing the regions of clothes of various types and studying the importance of such mid-level information in fashion parsing. Furthermore, our method can work on high-resolution images and does not need downsampling in contrast to these approaches. In clothing recognition, Hara *et al.* [36] combine convolutional neural network and object detection to find fashion items. Yamaguchi *et al.* [37] apply a similar method to recognise clothing attributes. Both of the approaches locate clothes by bounding boxes rather than pixel-wise labelling.

The work discussed in this paper aims at parsing fashion image into five mid-level semantic parts including *upper-clothing*, *lower-clothing*, *hair*, *skin* and *background*. The proposed method starts from super-pixels obtained by a popular over-segmentation method [38]. The key idea is to select some super-pixels which have high probabilities of being the predefined mid-level parts as seed regions. These seed regions are then expanded to be the full mid-level parts. Owing to the coherence of pose skeleton and these mid-level parts, super-pixels at specific positions on pose skeleton are set to be seeds. For example, super-pixels on legs are set to be the seed regions of *lower-clothing*. Then, more adjacent super-pixels are assembled by conditional inference according to the position of super-pixels and perceptual similarity between neighbour regions. The inference is implemented by CRFs model on a chain structure where nodes stand for super-pixels. The super-pixel chains are generated from expanding trees with a greedy criterion, where the seeds are used as tree roots. The probabilities of a super-pixel belonging to a certain part can be estimated under differently configured chains. To mix up the multiple probabilities to obtain the final parsing results, a voting stage is adopted. To present the importance of the mid-level parsing, a re-parsing method is also proposed. In this method, the mid-level parts are used to re-parse the labelling results output by high-level parsing algorithms. This process can correct the regional errors of clothes effectively.

The rest of this paper is organised as follows: in Section 2, the mid-level parsing method for fashion images is described in detail. In Section 3, the re-parsing process is presented. In Section 4, the experimental results of the mid-level parsing and re-parsing in high-level are shown. Finally, the conclusions are made in Section 5.

## 2 Parsing fashion image into mid-level parts

### 2.1 Parsing process

The pipeline of the proposed method is shown in Fig. 2. For a given fashion image $I$, the segmentation algorithm in [38] is applied to obtain the super-pixels $S_I = \{s_i\}$, $i \in [1, N]$ where $N$ is their amount. A human pose detector [23] is used to estimate the pose skeleton. The target is to label each super-pixel $s_i \in S_I$ with a mid-level part label in $T = \{upper\text{-}clothing, lower\text{-}clothing, skin, hair, background\}$. The proposed method labels each semantic part $t \in T$ at first. That is, for each $t \in T$, the first step is to label the image into two parts: $t$ (positive) and not-$t$ (negative). Previous studies [16, 18] model the spatial relationship of the labels directly with pairwise potential. In our approach, human pose is applied to give out the high probability locations of clothing parts and the spatial relationship among different labels is assured by the seed selection process with pose structure. Thus, we can assume that the labels are independent in spatial. Our method then parses one part while regarding all the others as 'negative' parts in each region expanding process in this stage.

When labelling a full mid-level part by region expanding, super-pixels crossed by the segments of upper body, lower body, head and full body are set to be seed regions of upper-clothing, lower-clothing, hair and foreground, respectively. Seeds for skin are not selected, because skin appears in many locations on human body and could be better detected by a fully trained classifier based on colour. Then, the expanding trees are generated, which indicate the possible expanding paths. Adjacent super-pixels with similar appearance at the boundary are possibly belonging to the same part. In this sense, for a given $t$, a greedy method is applied to generate the expanding trees from the seeds. Let $U$ be the set of super-pixels already on the expanding trees, which initially contains only seeds. Let $U'$ be the set of super-pixels adjacent to the ones in $U$. The greedy criterion is connecting each super-pixel in $U'$ to the one in $U$ with the weakest average gradient magnitude [39] at the boundary. Then, $U$ is updated by the operation

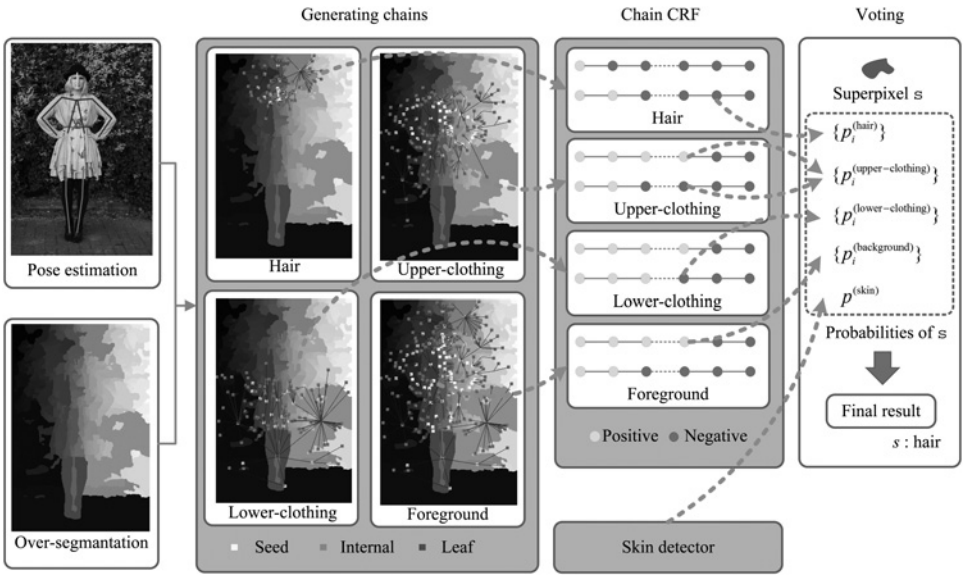


**Fig. 2** *Pipeline of the proposed mid-level parsing method*

$U = U' \cup U$. After that, $U'$ is rebuilt to include the super-pixels adjacent to the ones in the updated $U$. This process is iterated until no super-pixel is added to $U'$. It is apparent that super-pixels which are located at a certain distance away from the seeds are possibly backgrounds. To avoid unnecessary calculation in later processes, four masks are generated to indicate the coarse regions of upper-clothing, lower-clothing, hair and foreground (as shown in Fig. 3). Super-pixels in $U'$ are then limited by the overlapping rate computed with the corresponding mask.

*2.1.1 Chain-CRFs:* One possible way for inference is conducting CRFs model based on the expanding trees. However, the size and connectivity of the expanding trees still vary dramatically. It is hard to regularise the model parameters to be adaptive for all situations. For simplicity, it is assumed in this paper that the edges in the trees between one node and its parent node are the only feasible path for expanding to the node. Under this assumption, the super-pixel chains from roots to leaves are extracted, which compose a chain set $A_t = \{\alpha_j\}$. For any chain $\alpha_j \in A_t$, the labelling process is based on the maximum a posteriori (MAP) assignment under a conditional probability distribution $P(l|\alpha)$, where $l$ denotes the labels assigned to the nodes in $\alpha_j$

$$\hat{l} = \arg\max_{l} (P(l|\alpha_j)) \tag{1}$$

This conditional probability distribution is modelled by chain-CRFs. Accurate estimation of human pose can assure the selected seed super-pixels located in foreground region. Therefore, setting the label of the seeds to be 'positive' in the inference can avoid incorrect predictions on them. Unfortunately, the pose estimation sometimes fails in local joints of human body, especially on the arms. To reduce the impact of inaccurate poses, a unary feature based on the global statistics of colour distribution is proposed which will be discussed in Section 2.2.

*2.1.2 Global voting:* The inferences in the chain-CRFs take binary labels. The estimated probability of one node taking each label varies in different chain configurations. To solve the contradiction, a global voting process is applied for finding the most reasonable label in $T$ as the final result. This process is defined as computing the maximum of the average labelling

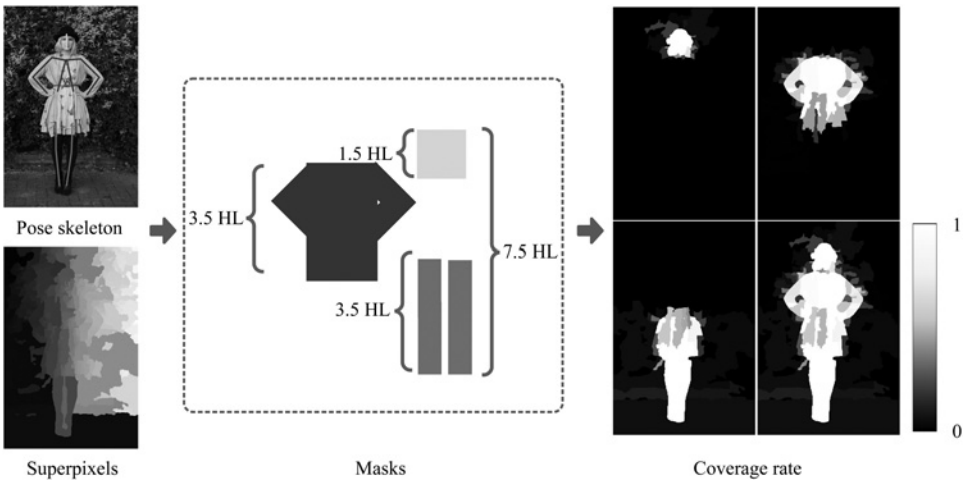


**Fig. 3** *Masks for head, upper body and lower body are shown. Scales are measured by HL. The combination of the three masks is used as a full body mask. These masks indicate the coarse regions of hair, upper-clothing, lower-clothing and foreground, respectively. The coverage rate of super-pixels with the four masks are shown in brightness*


*IET Image Process.*, 2016, Vol. 10, Iss. 6, pp. 456–463

probability for each super-pixel

$$\hat{l}_s = \arg\max_t \left( \frac{1}{M_{s,t}} \sum_{i=1}^{M_{s,t}} p_{s,i,t} \right), \quad t \in T \tag{2}$$

where $p_{s,i,t}$ is the estimated probability of super-pixel $s$ taking label $t$ in a single chain prediction. As described in previous section, the super-pixel chains are generated from different seeds according to the processing label. Thus, $M_{s,t}$ is applied to average the probabilities, which is the amount of predictions on super-pixel $s$ in the parsing process for $t$.

### 2.2 Conditional training and inference of chain-CRFs

The probability distribution $P(l|\alpha)$ of a chain $\alpha$ containing $N$ nodes in (1) is defined as

$$P(l|\alpha) = \frac{1}{Z(\alpha)} \prod_{i=1}^{N} \Phi(l^{(i)}, \phi_i) \prod_{(i,j)\in E} \Psi(l^{(i)}, l^{(j)}, \psi_{i,j}) \tag{3}$$

where $E$ is the set of neighbour nodes in $\alpha$ and $Z$ is the partition function. $\Phi$ and $\Psi$ are the unary potentials and pairwise potentials, respectively. $\phi_i$ represents the concatenated unary features of the $i$th node in $\alpha$ and $l^{(i)}$ is the assigned label. Similarly, $\psi_{i,j}$ is the concatenated pairwise features. Both unary and pairwise features will be described later.

Binary labels are taken in the chain-CRFs model. That is $l^{(i)} \in \{0, 1\}$ representing the two labels: $t$ and *not-t*. The unary potentials and pairwise potentials are modelled as follows

$$\ln \Phi(l^{(i)}, \phi_i) = \left( \sum_{k=1}^{n} \lambda_{0,k} \phi_i^{(k)}, \quad \sum_{k=1}^{n} \lambda_{1,k} \phi_i^{(k)} \right) \tag{4}$$

$$\ln \Psi_{i,j}(l^{(i)}, l^{(j)}|\psi_{i,j}) = \begin{pmatrix} \sum_{k=1}^{m} \theta_{0,0,k} \psi_{i,j}^{(k)} & \sum_{k=1}^{m} \theta_{0,1,k} \psi_{i,j}^{(k)} \\ \sum_{k=1}^{m} \theta_{1,0,k} \psi_{i,j}^{(k)} & \sum_{k=1}^{m} \theta_{1,1,k} \psi_{i,j}^{(k)} \end{pmatrix} \tag{5}$$

where $m$ and $n$ are the amount of unary features and pairwise features, respectively. $\lambda$ and $\theta$ are the parameters of the chain-CRFs model.

In the proposed model, unary features and pairwise features are organised as follows.

*2.2.1 Unary features:* Unary features are extracted from single super-pixel. Two kinds of unary feature are designed: coverage rate vector (CRV) and colour similarity to masks (CSM). CRV is calculated by utilising the proportion of human figure, which is defined based on a series of masks. These masks indicate the coarse regions of the target parts. To measure the proportion of human figure, head length (HL) is widely used as the basic unit [40, 41]. As shown in Fig. 3, four masks are generated to indicate the coarse regions of upper-clothing, lower-clothing, hair and foreground. HL is applied to set the scale of the masks. CRV is a vector where the values are the proportions of area covered by the four masks. CSM computes the correlation coefficients between the colour histograms calculated from a super-pixel region and the masked regions. These histograms are calculated in red-green-blue (RGB) colour space with 16 bins for each channel in this paper.

**Table 1** Mapping table for mid-level part labels and high-level labels

| Mid-level part labels | High-level labels |
|---|---|
| upper-clothing | coat, dress, shirt, blazer, sweater, … |
| lower-clothing | pants, skirt, jeans, shorts, … |
| hair | hair |
| skin | skin |
| background | background, bag, … |

*2.2.2 Pairwise features:* Pairwise features are extracted from adjacent super-pixels in the chain. Two kinds of pairwise features are extracted. The first one is the correlation coefficient of colour histograms from adjacent super-pixels. The second one is the gradient at the border of adjacent super-pixels. The recent edge detector [39] is applied to obtain the gradient. The mean value is used as the feature.

At last, limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS) [42] is applied to train the model parameters. Belief propagation is used to obtain MAP assignments. The chain-CRFs model is constructed with undirected graphical models (UGM) [43] implementation.

### 2.3 Skin detection

Most of the skin detectors [44–47] work on pixel level. In this paper, the per-pixel Naive Bayes skin detector [44] is extended to be applicable for super-pixels. First, a pixel-level Bayes classifier is trained as the probabilistic model of colour $c$ being skin $P(\mathbf{skin}|c)$ and being non-skin $P(\mathbf{non-skin}|c)$. The probabilities of a super-pixel $s$ being skin $P(\mathbf{skin}|s)$ and non-skin $P(\mathbf{non-skin}|s)$ are then defined as

$$P(\mathbf{skin}|s) = \frac{1}{N} \sum_{c\in s} P(\mathbf{skin}|c) \tag{6}$$

$$P(\mathbf{non-skin}|s) = \frac{1}{N} \sum_{c\in s} P(\mathbf{non-skin}|c) \tag{7}$$

where $c$ is the colour of a pixel in $s$ in RGB colour space and $N$ is the amount of pixels in $s$. The classification rule is defined with a constant threshold value $\eta$ as

$$\gamma = \frac{P(\mathbf{skin}|s)}{P(\mathbf{non-skin}|s)} > \eta \tag{8}$$

In practice, $\eta$ is set to be 3 in the experiments.

## 3 Application: clothing re-parsing

The mid-level parts can be used as the bridge between low-level super-pixels and high-level semantic labels. These parts can be applied to improve the performance of fashion parsing. The re-parsing process is presented as an application of the proposed mid-level parts in fashion parsing. This process aims at boosting the parsing accuracy by correcting the regional errors of clothing items.

The re-parsing process starts with a label mapping operation. A mapping table between mid-level parts and high-level labels is shown in Table 1. In Table 1, clothes located exactly on upper/lower human body are mapped to upper-/lower-clothing. Long style clothes, which cover both upper and lower body, are mapped to upper-clothing. In the re-parsing process, two kinds of mid-level parsing results are obtained: the result of mapping the high-level parsing results to mid-level labels and the result of the proposed method. Then, any super-pixel $s$ with contradictory prediction is set to be the missing region of some other item nearby. It is assumed that each super-pixel $s$ has similar appearance with its belonging part. In this sense, the expanding tree (described in Section 2.1) is applied to iteratively get the parent node starting from $s$. The iteration process terminates when reaching a super-pixel $s'$ with the same part label to $s$. At last, $s$ is re-labelled with the high-level semantic label of $s'$. The re-parsing experiment is shown in Section 4.4.

## 4 Experiments

In this section, the experimental results of the proposed method are shown at first. Then the re-parsing result is presented. The

**Table 2** Parsing performances on Fashionista and CCP

| Methods | Acc., % | F.g. Acc., % | Average precise, % | Average recall, % | Average IoU, % | Average F-1, % |
|---|---|---|---|---|---|---|
| *Fasionista* | | | | | | |
| baseline | 77.59 | – | 15.52 | 20.00 | – | – |
| DDN [33] | 89.91 | 63.39 | 73.56 | 61.95 | 49.89 | 62.40 |
| PCF [16] | 88.26 | 69.21 | 67.17 | 72.61 | 54.11 | 68.41 |
| Paper Doll [17] | 91.71 | 70.55 | 74.48 | 77.25 | 61.86 | 75.27 |
| Ours (Cond.) | 92.84 | **76.51** | 78.79 | 77.83 | 65.16 | 77.76 |
| Ours | **93.01** | 76.34 | **79.56** | **78.10** | **65.87** | **78.04** |
| *CCP* | | | | | | |
| baseline | 77.60 | – | 15.52 | 20.00 | 20.00 | – |
| DDN [33] | 87.68 | 60.63 | 69.91 | 61.13 | 48.96 | 62.31 |
| PCF [16] | 87.24 | 54.76 | 63.47 | 58.94 | 45.65 | 57.03 |
| Paper Doll [17] | 91.13 | 70.72 | 71.53 | **78.54** | 60.04 | 73.54 |
| Ours | **92.16** | **76.06** | **77.85** | 77.23 | **64.80** | **77.50** |

over-segmentation and pose estimation algorithms used in the experiments are the same as in [17].

### 4.1 Datasets

The experiments are conducted on two datasets: Fashionista [48] and Clothing Co-Parsing (CCP) [49]. Fashionista dataset contains 685 pixel-wise annotated fashion images with 56 high-level semantic labels. The training set and testing set contains 456 and 229 images, respectively. CCP is a newly proposed dataset by Yang *et al.* [19] for clothing co-parsing, containing 1004 pixel-level annotated fashion images with 57 high-level semantic labels. The clothes in the two datasets have a wide range of types and styles.

### 4.2 Evaluation metrics

The first comparison experiment is conducted on Fashionista dataset. To evaluate the effectiveness of the proposed approach, the parsing accuracy is compared with three state-of-the-art approaches: deep decompositional network (DDN) [33], PCF [16] and Paper Doll [17]. Though these approaches parse clothing in high-level, the results of mid-level parts can be obtained by label mapping according to Table 1.

Another experiment is conducted on CCP dataset to present the generalisation capability of the new method. That is, the model parameters are not re-trained when parsing on CCP, but identical to those trained on Fashionista. This is very challenging because of the more complicated and indistinct backgrounds in CCP.

The performance is measured in terms of standard metrics as in [17]: accuracy, foreground accuracy, average precise, average recall and average F-1 measure over pixels. The background bias problem is significant in fashion images. In the two datasets, naively predicting all pixels to be *background* will result in 77.59 and 77.60% pixel accuracy. The evaluation metric intersection-over-union (IoU) can address this problem [50]. Therefore, we add it in the experiments to make the comparison more sensible.

### 4.3 Experimental results

Table 2 summarises the performance of our method on mid-level parsing. That is labelling fashion images into the proposed five mid-level parts. The upper part of Table 2 summarises the performance of our method on Fashionista dataset compared with DDN [33], PCF [16] and Paper Doll [17]. Among the listed rows, 'Ours(Cond.)' represents the results of the proposed method based on the inference where seeds are set to be positive nodes. That is a hypothesis of the estimated pose is accurate and the seeds of each target part are correctly selected. 'Ours' shows the results of the proposed method. The comparison of 'Ours(Cond.)' and 'Ours' shows that the proposed method can reduce the impact of inaccurate human poses. To further illustrate the impact of pose accuracy, the proposed method is tested on ground truth pose skeleton annotated in Fashionista, which leads to a 93.84% overall accuracy. This is very similar to the result based on estimated pose. In this experiment, the proposed method outperforms the three state-of-the-art works on all standard metrics. The overall accuracy is boosted from 91.71 to 93.01%.

The result of the cross-dataset experiment is shown in the lower part of Table 2. The parsing performance of our method is very close to the results in Fashionista with 92.16% overall accuracy. The proposed method outperforms DDN and PCF on all standard metrics and outperforms Paper Doll on five metrics.

F-1 scores on Fashionista for all mid-level parts are shown in Fig. 4, compared with the method DDN [33], PCF [16] and Paper

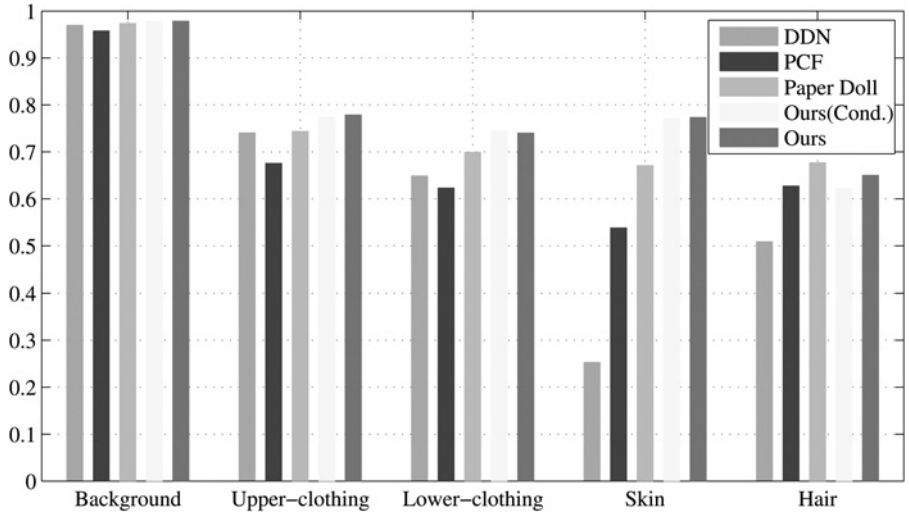


**Fig. 4** *F-1 scores for the mid-level parts on Fashionista*


*IET Image Process.*, 2016, Vol. 10, Iss. 6, pp. 456–463


**Fig. 5** *Some successfully parsed examples and comparison*

Doll [17]. The proposed method achieves superior performances over these methods on four of the five labels. Different from the other part labels, *hair* has much larger variation and some long hair is totally out of the given mask. This is the reason for our method obtaining a lower F-1 score. The mask for hair is not expanded because it will weaken the accuracy of clothing parts which are mainly concerned in this paper. Figs. 5 and 6 illustrate some successful and failed parsing results from Fashionista. Our method produces actual clothing boundary even in some challenging situations, for example, clothing with complex texture (as shown in the first column of Fig. 5) and various shapes such as vest and long coat in the second column of Fig. 5. However, it may get wrong results under some scenarios such as clothes with skin such as colour and extremely disordered background.

The proposed approach in this paper is light-weighted and effective in mid-level fashion parsing. DDN [33] only works on fixed-size images and their downsampling process potentially weakens the pixel-wise precision. However, the proposed method can process high-resolution fashion images and gives out exact boundaries of clothing items with various structures and appearances in the two datasets. As shown in Fig. 4, our method achieves the best performance on the regional accuracy of clothes. Our method does not rely on additional dataset as in [17], but exploits more useful information from human pose and prior physical structure of the human body such as the generated coarse masks for clothes. The regularised chain-CRFs model makes the region expanding process efficient and easier to assemble clothing parts than the conventional approaches [16]. Paper Doll [17] utilises nearest neighbours searched from a huge retrieval dataset. However, their method will not work well when parsing clothes whose pattern or style seldom appears in the retrieval dataset. Our method selects multiple seed super-pixels from high probability locations of the targets to overcome the inconsistency in the clothing area, which is not disturbed by auxiliary dataset. In the

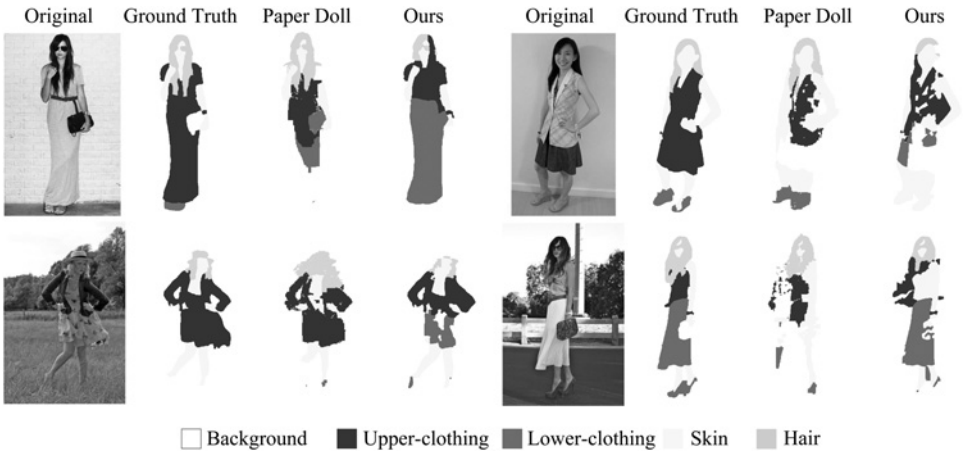


**Fig. 6** *Some failed example of the proposed method and comparison*

**Table 3** Performance of clothing re-parsing on Fashionista

| Methods | Acc., % | F.g. Acc., % | Average precise, % | Average recall, % | Average IoU, % | Average F-1, % |
|---|---|---|---|---|---|---|
| PCF | 77.45 | 23.11 | 10.53 | 17.20 | 5.46 | 10.35 |
| re-parsed PCF | 80.79 | 23.72 | 12.31 | **17.30** | 6.33 | 11.91 |
| Paper Doll | 84.68 | 40.20 | 33.34 | 15.35 | 8.03 | 14.87 |
| re-parsed Paper Doll | **86.34** | **40.67** | **36.31** | 14.97 | **8.30** | **15.33** |

experiment, the proposed approach obtains superior regional accuracy on parsing various clothing items.

The experiments are carried out on an Intel I7-4790 central processing unit at 3.6 GHz and 16 GB random access memory personal computer. Images in Fashionista and CCP are segmented into 200–400 super-pixels. Our method takes 2–3 s to parse a single image, while DDN, PCF and Paper Doll run in 0.16, 1520 and 20–40 s, respectively. The DDN method is much faster because of its strong limitation on the scale of the input image (about 1/50 of the original image). The proposed method is much faster than PCF and Paper Doll. Since, our method does not need to construct graph structure on all super-pixels as in PCF, but only on the super-pixels selected by the masks. Also, our method does not need a huge dataset for retrieving nearest neighbours such as Paper Doll, but parsing by exploiting appearance information within the image.

Though our method is successful on parsing the proposed mid-level parts, there are some drawbacks of the method. Our method presumes that at most one visible upper-clothing and one lower-clothing item appear in the target image. This is correct in most of the case. For example, in CCP dataset, 973 out of 1004 images are to be this case. However, there might still be two visible upper-clothing items on human body and our method labels them into one upper-clothing part. Owing to the large colour and texture variations, solving this issue with low-level features is very challenging. This question might be better solved by combining with attribute recognition in our future work. Some small items such as belts are ignored in our method. Since, this paper focuses on mid-level study for fashion parsing and the major issue in fashion parsing is dealing with the variation of clothes. These items are usually too small to have significant effect on the overall appearances.

### 4.4 Clothing re-parsing

The clothing re-parsing method is experimented based on the results of PCF [16] and Paper Doll [17] on Fashionista with all 56 high-level


**Fig. 7** *Examples of clothing re-parsing based on the results of PCF and Paper Doll*


*IET Image Process.*, 2016, Vol. 10, Iss. 6, pp. 456–463

labels. After re-parsing, the overall pixel accuracy of PCF is improved from 77.45 to 80.79% and the accuracy of Paper Doll is improved from 84.68 to 86.34%. The results are listed in Table 3.

The proposed re-parsing approach aims at fixing the regional errors of clothing items in high-level parsing work. This experiment presents one possible way of applying the mid-level parts for improving the performance of fashion parsing in high-level. It can be seen in Fig. 7 that the boundaries of clothes become much more accurate after the re-parsing process. However, foreground accuracy is not improved much, because it is determined by the accuracy of both segmentation and prediction. Improvement in regional accuracy is weakened by incorrect prediction. This can be observed from the low IoU and F-1 score in Table 3, where the corresponding scores in Table 2 can be regarded as the upper boundaries. The basis parsing methods, PCF and Paper Doll in this experiment, predict the clothing categories based on local regions provided by super-pixels. It can be observed from the comparison that appearance information provided by super-pixels is insufficient for recognition. This is also the reason for us to propose the mid-level parsing work to obtain the regions of clothing items.

## 5 Conclusions and future work

This paper has proposed a new method for parsing fashion image into mid-level semantic parts based on a simple chain-CRFs model. We have also presented a re-parsing scheme to show the importance of our mid-level work in high-level parsing approaches. Experiments show that the proposed method outperforms three state-of-the-art parsing methods in regional accuracy under all evaluation metrics on Fashionista, and the re-parsing scheme corrects the regional errors of clothes produced by two high-level parsing approaches remarkably.

There are several possible aspects to refine our mid-level work in the future. Currently, our model is trained on images with strong pixel-level annotations. Weak supervisions such as in the form of bounding boxes and image-level annotations can only give out rough clothes regions. We plan to adjust the method to work with weakly annotated fashion images. In this problem, robust feature extraction and model training methods may be the key issues which need to be studied. Further investigating the underlying role of mid-level parsing in high-level tasks such as clothing recognition based on the clothing parts is another future work of us.

## 6 References


1 Hu, Z., Yan, H., Lin, X.: 'Clothing segmentation using foreground and background estimation based on the constrained Delaunay triangulation', *Pattern Recognit..*, 2008, **41**, (5), pp. 1581–1592
2 Hasan, B., Hogg, D.: 'Segmentation using deformable spatial priors with application to clothing'. British Machine Vision Conf., 2010, pp. 83.1–83.11
3 Wang, N., Ai, H.: 'Who blocks who: simultaneous clothing segmentation for grouping images'. Int. Conf. Computer Vision, November 2011, pp. 1535–1542
4 Freire-Obregon, D., Castrillon-Santana, M., Ramon-Balmaseda, E., *et al.*: 'Automatic clothes segmentation for soft biometrics'. Proc. IEEE Int. Conf. Image Processing, October 2014, pp. 4972–4976
5 Chen, H., Gallagher, A., Girod, B.: 'Describing clothing by semantic attributes'. European Conf. on Computer Vision, 2012, pp. 609–623
6 Shen, J., Liu, G., Chen, J., *et al.*: 'Unified structured learning for simultaneous human pose estimation and garment attribute classification', *IEEE Trans. Image Process.*, 2014, **23**, (11), pp. 4786–4798
7 Deng, Y., Luo, P., Loy, C.C., *et al.*: 'Learning to recognize pedestrian attribute', arXiv preprint arXiv:1501.00901, 2015
8 Si, L., Zheng, S., Guangcan, L., *et al.*: 'Street-to-shop: cross-scenario clothing retrieval via parts alignment and auxiliary set'. IEEE Conf. on Computer Vision and Pattern Recognition, June 2012, pp. 3330–3337
9 Wang, X., Zhang, T.: 'Clothes search in consumer photos via color matching and attribute learning'. Int. Conf. on Multimedia, 2011, pp. 1353–1356
10 Liu, S., Feng, J., Song, Z., *et al.*: 'Hi, magic closet, tell me what to wear!'. Int. Conf. on Multimedia, 2012, pp. 619–628
11 Kalantidis, Y., Kennedy, L., Li, L.J.: 'Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos'. Int. Conf. Multimedia Retrieval, 2013, pp. 105–112
12 Gallagher, A.C., Chen, T.: 'Clothing cosegmentation for recognizing people'. IEEE Conf. on Computer Vision and Pattern Recognition, June 2008, pp. 1–8
13 Anguelov, D., Lee, K.C., Gokturk, S.B., *et al.*: 'Contextual identity recognition in personal photo albums'. IEEE Conf. on Computer Vision and Pattern Recognition, June 2007, pp. 1–7
14 Lin, D., Kapoor, A., Hua, G., *et al.*: 'Joint people, event, and location recognition in personal photo collections using cross-domain context'. European Conf. on Computer Vision, 2010, pp. 243–256
15 Yang, M., Yu, K.: 'Real-time clothing recognition in surveillance videos'. Int. Conf. Image Processing, September 2011, pp. 2937–2940
16 Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., *et al.*: 'Parsing clothing in fashion photographs'. IEEE Conf. on Computer Vision and Pattern Recognition, June 2012, pp. 3570–3577
17 Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., *et al.*: 'Retrieving similar styles to parse clothing', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, **37**, (5), pp. 1028–1040
18 Jammalamadaka, N., Minocha, A., Singh, D., *et al.*: 'Parsing clothes in unrestricted images'. British Machine Vision Conf., 2013, pp. 88.1–88.11
19 Yang, W., Luo, P., Lin, L.: 'Clothing co-parsing by joint image segmentation and labeling'. IEEE Conf. on Computer Vision and Pattern Recognition, June 2014, pp. 3182–3189
20 Farabet, C., Couprie, C., Najman, L., *et al.*: 'Learning hierarchical features for scene labeling', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, (8), pp. 1915–1929
21 Pinheiro, P., Collobert, R.: 'Recurrent convolutional neural networks for scene labeling'. Int. Conf. on Machine Learning, June 2014, vol. 32, pp. 82–90
22 Gupta, S., Girshick, R., Arbeláez, P., *et al.*: 'Learning rich features from RGB-D images for object detection and segmentation'. European Conf. on Computer Vision, 2014, pp. 1–16
23 Yang, Y., Ramanan, D.: 'Articulated pose estimation with flexible mixtures-of-parts'. IEEE Conf. on Computer Vision and Pattern Recognition, June 2011, pp. 1385–1392
24 Sapp, B., Taskar, B.: 'MODEC: multimodal decomposable models for human pose estimation'. IEEE Conf. on Computer Vision and Pattern Recognition, June 2013, pp. 3674–3681
25 Toshev, A., Szegedy, C.: 'DeepPose: human pose estimation via deep neural networks'. IEEE Conf. on Computer Vision and Pattern Recognition, June 2014, pp. 1653–1660
26 Liu, C., Yuen, J., Torralba, A.: 'Nonparametric scene parsing via label transfer', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, **33**, (12), pp. 2368–2382
27 Girshick, R., Donahue, J., Darrell, T., *et al.*: 'Rich feature hierarchies for accurate object detection and semantic segmentation'. IEEE Conf. on Computer Vision and Pattern Recognition, 2014, pp. 2–9
28 Gould, S., Rodgers, J., Cohen, D., *et al.*: 'Multi-class segmentation with relative location prior', *Int. J. Comput. Vis.*, 2008, **80**, (3), pp. 300–316
29 Cho, M.S., Seok, J.H., Lee, S., *et al.*: 'Scene text extraction by superpixel CRFs combining multiple character features'. Int. Conf. on Document Analysis and Recognition, September 2011, pp. 1034–1038
30 Ramanan, D.: 'Learning to parse images of articulated bodies', *Adv. Neural Inf. Process. Syst.*, 2007, **19**, pp. 1129–1136
31 Zhang, H., Wang, J., Tan, P., *et al.*: 'Learning CRFs for image parsing with adaptive subgradient descent'. Int. Conf. on Computer Vision, 2013, pp. 3080–3087
32 Sutton, C., McCallum, A.: 'Piecewise training for undirected models'. Annu. Conf. on Uncertainty in Artificial Intelligence, 2005, pp. 568–575
33 Luo, P., Wang, X., Tang, X.: 'Pedestrian parsing via deep decompositional network'. Int. Conf. on Computer Vision, December 2013, pp. 2648–2655
34 Liang, X., Liu, S., Shen, X., *et al.*: 'Deep human parsing with active template regression', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, **37**, (12), pp. 2402–2414
35 Liu, S., Liang, X., Liu, L., *et al.*: 'Matching-cnn meets knn: quasi-parametric human parsing'. IEEE Conf. on Computer Vision and Pattern Recognition, 2015
36 Hara, K., Jagadeesh, V., Piramuthu, R.: 'Fashion apparel detection: the role of deep convolutional neural network and pose-dependent priors'. IEEE Winter Conf. on Applications of Computer Vision, 2014
37 Yamaguchi, K., Okatani, T., Sudo, K., *et al.*: 'Mix and match: joint model for clothing and attribute recognition'. British Machine Vision Conf., 2015
38 Felzenszwalb, P.F., Huttenlocher, D.P.: 'Efficient graph-based image segmentation', *Int. J. Comput. Vis.*, 2004, **59**, (2), pp. 167–181
39 Dollar, P., Zitnick, C.L.: 'Structured forests for fast edge detection'. Int. Conf. on Computer Vision, 2013, pp. 1841–1848
40 Jadav, H., Shah, G.: 'Determination of personal height from the length of head in Gujarat region', *J. Anatomical Soc. India*, 2004, **53**, (1), pp. 20–21
41 Sudhir, P., Zambare, B., Shinde, S., *et al.*: 'Determination of personal height from the length of head in Maharashtra region', *Indian J. Forensic Med., Pathol.*, 2010, **3**, (2), pp. 55–58
42 Liu, D.C., Nocedal, J.: 'On the limited memory BFGS method for large scale optimization', *Math. Program.*, 1989, **45**, (1–3), pp. 503–528
43 UGM. Available at 'http://www.cs.ubc.ca/schmidtm/software/ugm.html, accessed 2007
44 Jones, M.J., Rehg, J.M.: 'Statistical color models with application to skin detection', *Int. J. Comput. Vis.*, 2002, **46**, (1), pp. 81–96
45 Vezhnevets, V., Sazonov, V., Andreeva, A.: 'A survey on pixel-based skin color detection techniques', *Proc. Graphicon*, 2003, **3**, pp. 85–92
46 Phung, S.L., Bouzerdoum, A., Chai, D.: 'Skin segmentation using color pixel classification: analysis and comparison', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (1), pp. 148–154
47 Xu, T., Wang, Y., Zhang, Z.: 'Pixel-wise skin colour detection based on flexible neural tree', *IET Image Process.*, 2013, **7**, (8), pp. 751–761
48 Fashionista. Available at 'http://www.vision.is.tohoku.ac.jp/kyamagu/research/paperdoll/, accessed 2014
49 CCP. Available at 'https://www.code.google.com/p/clothing-co-parsing/, accessed 2014
50 Everingham, M., Van Gool, L., Williams, C.K.I., *et al.*: 'The PASCAL visual object classes (voc) challenge', *Int. J. Comput. Vis.*, 2010, **88**, (2), pp. 303–338