# Parsing Clothes in Unrestricted Images

Nataraj Jammalamadaka
nataraj.j@research.iiit.ac.in
Ayush Minocha
ayush.minocha@students.iiit.ac.in
Digvijay Singh
digvijay.singh@students.iiit.ac.in
C. V. Jawahar
jawahar@iiit.ac.in

Center for Visual Information Technology
IIIT Hyderabad
India, 500032

**Abstract**

Parsing for clothes in images and videos is a critical step towards understanding the human appearance. In this work, we propose a method to segment clothes in settings where there is no restriction on number and type of clothes, pose of the person, viewing angle, occlusion and number of people. This is a challenging task as clothes, even of the same category, have large variations in color and texture. The presence of human joints is the best indicator for cloth types as most of the clothes are consistently worn around the joints. We incorporate the human joint prior by estimating the body joint distributions using the detectors and learning the cloth-joint co-occurrences of different cloth types with respect to body joints. The cloth-joint and cloth-cloth co-occurrences are used as a part of the conditional random field framework to segment the image into different clothing. Our results indicate that we have outperformed the recent attempt [16] on H3D [3], a fairly complex dataset.

## 1 Introduction

Cloth parsing involves locating and describing all the clothes (*e.g.*, T-shirt, shorts) and accessories (*e.g.*, bag) that the person is wearing. As it describes the human appearance, it is a next step to human detection and pose estimation in understanding the images. It plays an important role in human pose estimation [16], action recognition, person search [7, 15], surveillance, cloth retrieval [9] and has applications in fashion industry [16]. Commercially, it can be used in online cloth retail portals where people can try out various clothes. The main challenges in solving this include the large variety of clothing patterns that have been developed across the globe by different cultures. Even within the same cloth type, say T-shirt, there is a significant variation in color, texture and other complicated patterns. Occlusions from other humans or objects, viewing angle and heavy clutter in the background further complicates the problem.

In the recent past, there has been considerable attention given to understanding and modelling clothes. Many methods [5, 8, 12, 14, 17] have been proposed to segment clothes in restrictive settings. In [17], cloth recognition is proposed for surveillance videos where the

camera angles are fixed and background subtraction can be effectively used for human detection. In [5], a real time upper body cloth segmentation is proposed in images where people are wearing a monochromatic clothing and printed/stitched textures. In [8], the method models a particular cloth combination accounting for color and shape. In [12], only the regions around the face like skin, hair and background are segmented. In [14], multi-person clothing segmentation algorithm is proposed. Given an image with multiple people occluding each other, the method is restricted to segmenting the upper cloths.

Our work is closest to that of Yamaguchi *et al*. [16]. They have proposed a method to parse clothes in fashion photographs. In their work, cloth parsing problem is viewed as an object segmentation using CRFs. While using segmentation algorithms is quite standard and has been applied earlier [13], their innovation comes in defining the unary potential. They use a pose estimation algorithm [18] to model a cloth type. Unlike our work, Yamaguchi *et al*. [16] have limited themselves to restricted settings where a single person is standing upright against a relatively simple background. Real images on the other hand are far more complex containing multiple people present against a complicated background and potentially occluded.

In our work, we aim to segment clothes in unconstrained settings. We handle the the diverse set of challenges by modelling the cloth appearance and its vicinity to a body part. While human pose estimation algorithms [2, 6, 11, 17] provide the body part configuration, they frequently fail and give wrong pose estimates when there are large occlusions and heavy clutter. To handle these challenges, it is more desirable to have a probabilistic estimate of the body part position than to have single, deterministic but a wrong pose estimate. Poselets [3] offers one such flexibility where they detect a combination of body parts. In this work we adapt poselets to locate human joints and model the cloth-joint co-occurrences by learning a function that assigns a cloth label based on the relative location with respect to joints. Since the neighboring regions which have similar appearance share the same label, we use the conditional random fields to segment different cloth labels in the image.

The remainder of the paper is organized as follows. Section 2 describes how robust human joints can be obtained under heavy clutter and occlusion. Section 3 describes the proposed approach. Section 4 discusses the interesting patterns that can be mined on a dataset. Finally section 5 presents the experimental results.

## 2 Robust human joint estimation

For modelling the clothes in unrestricted settings a robust human joint estimator is needed as clothes are worn around human joints. The popular choice to obtain human body part configuration $Z$ are human pose estimation algorithms [2, 6, 11, 18]. Unfortunately the standard pose estimation algorithms fail to detect occlusions. This is mainly because the top-down model of pose estimation algorithms cannot model occlusions effectively. Figure 1 displays the output of pose estimation algorithm [18] on images with large occlusions.

We therefore employ poselets [3] which are human body part detectors. Poselet models a combination of human body joints (*e.g*., face, shoulders and background in row 1 of figure 2). A particular combination is chosen based on how frequently it occurs. It is then trained using the support vector machines. Unlike pose estimation algorithms, poselets are immune to occlusion and clutter. Since they are not constrained by top-down geometric constraints, they do not assume the presence of all body parts. Poselets corresponding to a missing body part will simply not fire. In all, 150 poselets which cover different parts of

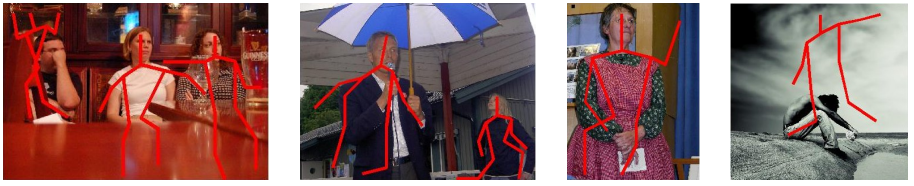

Figure 1: Pose Estimates [18] on H3D dataset: In the images containing severe occlusions and clutter, pose estimation fails.

the body are trained [5]. Given an image, each poselet locates the corresponding body part by giving multiple bounding boxes (termed poselets as well). In our implementation, we set the maximum number of poselet clusters to 10. The poselets detected in an image are validated by mutual co-occurrence and all inconsistent poselets are discarded [4]. Using these poselets, the torso and the bounding box of the person are estimated.

Although poselets coarsely locate body parts in the form of bounding boxes, they do not give the exact human joint locations. We solve this problem by annotating the relevant body joint locations in each of the 150 poselets. For example in a poselet modelling the face and shoulder, forehead, chin, neck and the shoulders are manually marked in a normalized space. Given an instance of this poselet in an image, these annotated points are appropriately scaled and translated to get the body joint locations modelled by the poselet. In this paper, we consider 17 points *viz.*, background, Head, Neck, and two each of torso, shoulders, elbows, wrists, hips, knee and Ankle. Our algorithm takes poselets, torso, bounding box of the person and the body joint locations as input.

## 3 Cloth parsing method

We model cloth parsing as a segmentation problem and assign each image region a label from the set $L$. Super-pixels are used as the basic image regions on the assumption that pixels which are adjacent and have similar appearance share the same label. Furthermore it can be observed that neighboring image regions have correlation in labels and it is certainly true in case of clothes. Thus Markovian assumption in space is valid and we use conditional random fields (CRF), represented by the undirected graph $G = (V,E)$, to model the cloth parsing problem.

Given an image, first the superpixels and body joint locations are computed. These superpixels form the vertices $V$ of the CRF. Two superpixels which share a border are considered adjacent and are connected by an edge $e \in E$. The segmentation is obtained by taking the configuration with the maximum a posteriori probability, popularly called as MAP configuration. The best labeling using the CRF model is given by the equation,

$$\hat{L} = argmax_L P(L|Z,I), \tag{1}$$

where $L$ is the label set, $Z$ is a distribution of the body joint locations and $I$ is the image.

The MAP configuration of CRF probability function given by the equation 1 is computationally expensive to compute and is usually a NP-hard problem. We thus make a simplifying assumption that at most two vertices in the graph form a clique thus limiting the order of a potential to two. Thus the CRF factorizes into unary and pair-wise functions and the log

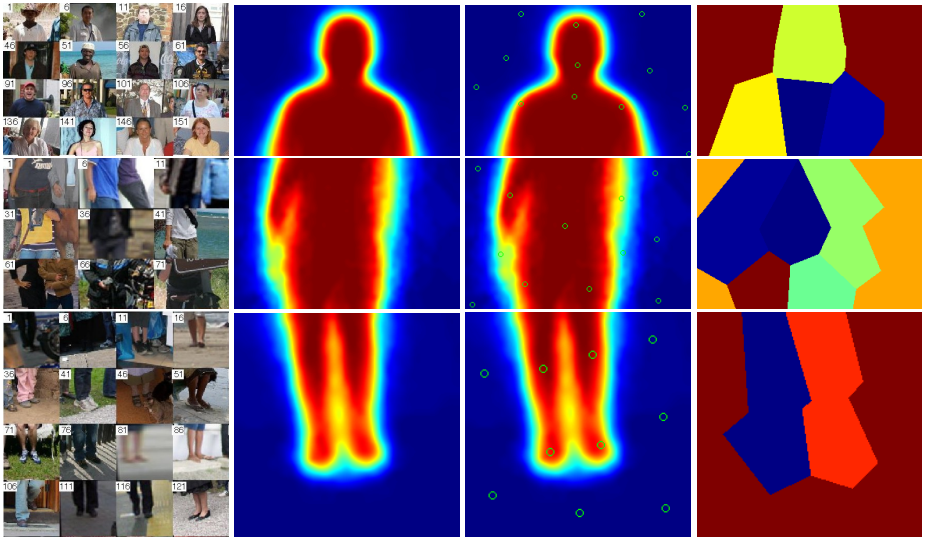

Figure 2: First, all the image regions in the training set which have a particular body configuration in common are selected (column one). Poselet [3] classifier is trained using these image regions to obtain a model and a mask (column two). In these poselet masks (column two), blue region represents the background and other color codes represent a human. We annotate the masks with body joints and background points (column three). Each of these masks exert an influence on an area around them (column four).

probability function is given by,

$$\ln P(L|Z,I) \equiv \sum_{i \in V} \Phi(l_i|Z,I) + \lambda_1 \sum_{(i,j) \in E} \Psi_1(l_i,l_j) + \lambda_2 \sum_{(i,j) \in E} \Psi_2(l_i,l_j|Z,I) - \ln G, \quad (2)$$

where $V$ is the set of nodes in the graph, $E$ is the set of neighboring pairs of superpixels, and $G$ is the partition function.

## 3.1 Unary potential

In CRFs, it is crucial to model the unary potential well for better performance. The unary potential function $\Phi$ models the likelihood of a superpixel $s_i$ taking the label $l_i$. First, using the estimated pose $Z = (z_1, ..., z_P)$ and the superpixel $s_i$, a feature vector $\phi(s_i, Z)$ is computed. Using the pre-trained classifier $\Phi(l_i|\phi(s_i,Z)) = \Phi(l_i|Z,I)$ for label $l_i$, a score is computed.

For the construction of features, the human body joint information, torso, bounding box and poselets are obtained using the procedure described in section 2. The feature vector $\phi(s_i,Z)$ for each superpixel consists of following histograms: (1) RGB color histogram, (2) CIE L*a*b color histogram, (2) histogram of Gabor filter responses, (4) histograms of normalized $X$ and $Y$ coordinates, and (5) histograms of normalized $X$ and $Y$ coordinates relative to each body joint location $z_p$. The $X$ and $Y$ coordinates are normalized by the width and height respectively of the bounding box containing the person.

The construction of the first four feature types is straightforward. The fifth feature type is constructed using the human body joint information $Z$ as follows. For all the super-pixels

which do not intersect with any of the poselet bounding boxes, the relative locations with respect to all the body joints are assigned infinity. For the super-pixel which intersects with a poselet bounding box, relative locations with respect to only the body joints present within the poselet bounding boxes are taken. For all body joints not present in the intersecting poselets, the relative location is assigned infinity. In case a super-pixel intersects with multiple bounding boxes, the relative position of each pixel with respect to a common part $z_p$ is weighed by the poselet score. Intuitively, this procedure can be understood as noting the relative position from a mean body joint location averaged using the poselet scores. Once the relative locations for all the pixels in a super-pixel are noted, a histogram is built.

### 3.2 Pair-wise potential

For the pairwise potential, we use the definitions from [16]. Pairwise potential, defined between two neighboring super-pixels, models the interaction between them. The pair-wise potential is defined in equation 2 as sum of two functions (called factors) $\Psi(l_i,l_j)$ and $\Psi(l_i,l_j|Z,I)$. The pairwise potential function $\Psi_1$ models the likelihood of two labels $l_i,l_j$ being adjacent to each other and $\Psi_2$ models the likelihood of two neighboring sites $s_i,s_j$ taking the same label given by the features $\phi(s_i,Z)$ and $\phi(s_j,Z)$ respectively. The function $\Psi_1$ is simply a log empirical distribution and $\Psi_2$ is a model learnt over all the label pairs respectively. The pairwise potential functions are given by,

$$\Psi_1(l_i,l_j),\ \ \Psi_2(l_i,l_j|Z,I) \equiv \Psi_2(l_i,l_j|\psi(s_i,s_j,Z)) \tag{3}$$

where $\psi(s_i,s_j,Z)$ is defined as,

$$\psi(s_i,s_j,Z) \equiv [(\phi(s_i,Z)+\phi(s_j,Z))/2, |(\phi(s_i,Z)-\phi(s_j,Z))/2|]\,. \tag{4}$$

### 3.3 Training and inference

Given a dataset containing images and cloth labels which include background, we wish to learn the cloth parser model. First for each data sample, superpixels and poselets are computed. All the superpixels which share a border are noted as neighbours. For each superpixel which falls within a poselet, the relative distance from the body joints present in the poselets are noted. At each superpixel the feature vector $\phi(s_i,Z)$ consisting of , (1) RGB color histogram, (2) CIE L*a*b color histogram, (2) histogram of Gabor filter responses, (4) normalized 2D coordinates within the image frame, and (5) normalized 2D coordinates with respect to each body joint location $z_p$ are noted. The bin size for the histograms is 10. Logistic regression is used to learn $\Phi(l_i|Z,I)$ and $\Psi_2(l_i=l_j|\psi(s_i,s_j,Z))$. Given a new image, the super-pixels, poselets and feature vector $\phi$ are computed. For each super-pixel, the unary potential and pairwise potential values are computed using the feature vector and the learnt models. The best label is inferred using the belief propagation implemented in libDAI [10] package. The parameters $\lambda_1,\lambda_2$ in the equation 2 are found by cross validation.

## 4 Clothing pattern Mining

Using the cloth labelling obtained from the algorithm described in the previous section, several interesting clothing patterns can be mined from a collection of images. In this section

two such interesting patterns *viz.*, cloth co-occurrences and upper and lower body cloth's color co-occurrences are explored on the Fashionista dataset.

The objective is to ascertain how many times does $(item_1, ..., item_n)$ item set co-occur. The classic apriori algorithm [1] is an efficient way to determine the co-occurrence frequency. A threshold $mt \in [0, 100]$, called minimum support, has to be specified by the user to prune low frequency item set co-occurrences. The algorithm uses the apriori property which states that any subset of an item set should have a minimum support. Initially frequent item sets of size one are generated that are above the minimum support. In the next step, frequent item sets of size two are generated by taking all the combinations from item sets of size one and then pruned based on minimum support. For the item sets of size greater than two, only those survive whose subsets have a minimum support. Finally the algorithm outputs the item set co-occurrences and their support values.

*Cloth co-occurrences* are obtained by applying the apriori algorithm on the cloth labels determined by the proposed method. To get interesting co-occurrences, frequent occurring labels like skin, hair *etc.*, have been removed. *Color co-occurrences* of upper-body and lower-body clothes are mined using the following procedure. First, a representative cloth type for the upper-body and lower-body are determined. This is done by selecting the outermost cloth worn by the person. For example, blazer is worn over t-shirt and hence it represents the upper-body cloth. Images that do not have either upper-body cloth or lower-body cloth labels specified above are ignored. The upper-body and lower-body clothes are then assigned a color as follows. The dominant RGB value in the upper-body or lower-body cloth image region is taken as its representation. This RGB value is then vector quantized to one of the 500 cluster centers. Using the map between colors and cluster centers, the image region of upper-body or lower-body is assigned one of blue, brown, orange, pink, red, violet, green and yellow. The apriori algorithm is then applied to obtain the frequent upper-body and lower-body color co-occurrences.

# 5 Experiments

We evaluate our method on two datasets, a) Fashionista and b) H3d. For each image in the above datasets, we compute the superpixels and poselets (described in section 2), both of which have standard implementations available on the internet. We then compute the cloth labelling using the method described in section 3. Using the ground truth segmentation masks, we evaluate our algorithm and also compare it with Yamaguchi *et al.* [16]. Two measures, pixel accuracy and mean average garment recall (mAGR) are used to quantitatively evaluate the algorithms. Pixel accuracy is total number of correct labelling in the image. It is a gross measure and is biased towards labels with large areas (*e.g.*, background). The measure, mAGR, is much more balanced measure and gives equal importance to labels of all the sizes. Two sets of parameters $\lambda_1, \lambda_2$, each are optimized for pixel accuracy and mAGR using cross validation. The outputs from these two parameters are termed *Full-a* and *Full-m* respectively. As a baseline, all the pixels are labelled as background and the above two measures are calculated. In the next two sections, the two datasets and results are described.

## 5.1 Datasets

**H3D:** This dataset has been introduced in [3]. It has a total of 180 images for training, 40 for validation and 107 images for testing. This dataset is derived from flickr images and is

| Method | Full-a | | Full-m | | Unary | |
|---|---|---|---|---|---|---|
| | Pixel acc | mAGR | Pixel acc | mAGR | Pixel acc | mAGR |
| [16] | 61.0 ± 5.0 | 34.9 ± 3.9 | 49.9 ± 4.5 | 39.9 ± 4.8 | 49.5 ± 4.3 | 39.6 ± 4.4 |
| Ours | 74.5 ± 4.7 | 49.7 ± 3.8 | 68.5 ± 5.4 | 55.2 ± 4.5 | 68.4 ± 5.4 | 54.8 ± 4.3 |
| Ours+Noc | 77.4 ± 4.0 | 57.0 ± 3.3 | 70.2 ± 5.2 | 63.1 ± 4.5 | 70.1 ± 5.2 | 62.4 ± 4.3 |

Table 1: Results on H3D: The baseline for accuracy is 74.7 ± 5.6 and for mAGR is 14.3 ± 0.6. 'Ours+Noc' indicates that the algorithm has been run with out the 'occluding object' label.

| Garment | [16] | Ours | Garment | [16] | Ours |
|---|---|---|---|---|---|
| background | 54.8 ± 4.6 | 74.0 ± 5.7 | shoes | 33.3 ± 11.7 | 51.0 ± 15.2 |
| upperclothes | 25.6 ± 6.0 | 65.6 ± 7.6 | bag | 12.8 ± 10.9 | 19.8 ± 12.0 |
| lowerclothes | 40.4 ± 9.4 | 59.9 ± 14.9 | occ-object | 13.5 ± 8.7 | 13.5 ± 5.8 |
| skin | 71.5 ± 10.5 | 62.0 ± 8.9 | hat | 19.2 ± 26.9 | 40.2 ± 21.6 |
| hair | 43.3 ± 11.0 | 62.2 ± 11.6 | socks | 0.0 ± 0.0 | 0.8 ± 2.7 |
| dress | 0.6 ± 1.3 | 0.1 ± 0.2 | sunglasses | 1.3 ± 4.1 | 13.6 ± 24.8 |

Table 2: Recall for selected garments on H3D.

very complex with severe occlusions and heavy clutter. The dataset has a total of 22 labels which include, face, hair, upperbody clothes, lowerbody clothes, hair, face, neck, left/right arm, left/right leg, left/right shoe, occluding object, bag, hat, dress, left/right glove, left/right sock and sunglasses. Since the main concern is cloth segmentation, the left/right part of the same cloth type is not relevant. All the left/right labels are thus converted into a single label (*e.g.*, left/right shoes → shoes). Furthermore, labelling different body parts like left/right leg, left/right arm, neck and face are not relevant and have been converted to a single label 'skin' in the image. Finally in case of occlusion from a person, the labels of the occluding person is considered and occlusion from an object is labelled as 'occluding object'. In all a total of 13 labels are present *viz.*, upperbody clothes, lowerbody clothes, occluding object, skin, hair, dress, shoes, bag, hat, socks, background and sunglasses.

**Fashionista:** This dataset has been introduced in [16]. It has a total of 685 images collected from Chictopia.com website made for fashion bloggers. Since these are fashion images, it includes a wide range of garment types and accessories. Each image has one person standing in an upright position with a clean background and is annotated with labels at a pixel level. In all there are about 56 labels. Broadly the labels can be classified as upper-body clothes, lower-body clothes, accessories, skin and hair.

## 5.2 Results

**H3d:** In this dataset, both the algorithms assign a label to each pixel. Since Yamaguchi *et al.* [16] assumes that there is a single person per image, we adapt it to multi-person images. First the pose estimation algorithm [18] available at the author's site is used to predict multiple pose estimates. Then features corresponding to the absolute position and pose-based ones are calculated relative to the bounding box defined by the pose estimate. In our opinion, we have made all the efforts to adapt the algorithm to make a fair comparison. Table 1 clearly indicates that our method outperforms [16] by about 13.5% and 15.3% in pixel accuracy and mAGR measures respectively. A similar trend can be seen in recall of individual labels in table 2. We also observed that when the 'occluding object' label is removed from the dataset

| Method | Full-a | | Full-m | | Unary | |
|---|---|---|---|---|---|---|
| | Pixel acc | mAGR | Pixel acc | mAGR | Pixel acc | mAGR |
| [16] | $89.0 \pm 0.8$ | $63.4 \pm 1.5$ | $88.3 \pm 0.8$ | $69.6 \pm 1.7$ | $88.2 \pm 0.8$ | $69.8 \pm 1.8$ |
| Ours | $87.7 \pm 0.8$ | $62.7 \pm 1.8$ | $86.0 \pm 1.0$ | $70.2 \pm 2.0$ | $85.9 \pm 1.1$ | $70.4 \pm 2.0$ |

Table 3: Results on Fashionista: The baseline for accuracy is $77.6 \pm 0.6$ and for mAGR it is $12.8 \pm 0.2$.

| Garment | [16] | Ours | Garment | [16] | Ours |
|---|---|---|---|---|---|
| background | $95.3 \pm 0.4$ | $92.5 \pm 0.9$ | jacket | $51.8 \pm 15.2$ | $49.4 \pm 15.5$ |
| skin | $74.6 \pm 2.7$ | $74.7 \pm 2.1$ | coat | $30.8 \pm 10.4$ | $29.5 \pm 12.5$ |
| hair | $76.5 \pm 4.0$ | $78.2 \pm 4.0$ | shirt | $60.3 \pm 18.7$ | $65.5 \pm 21.9$ |
| dress | $65.8 \pm 7.7$ | $59.2 \pm 10.4$ | cardigan | $39.4 \pm 9.5$ | $46.8 \pm 11.6$ |
| bag | $44.9 \pm 0.8$ | $44.0 \pm 7.6$ | blazer | $51.8 \pm 11.2$ | $54.7 \pm 9.5$ |
| blouse | $63.6 \pm 9.5$ | $63.1 \pm 10.3$ | t-shirt | $63.7 \pm 14.0$ | $68.2 \pm 10.7$ |
| shoes | $82.6 \pm 7.2$ | $80.3 \pm 9.5$ | socks | $67.4 \pm 16.1$ | $58.3 \pm 23.2$ |
| top | $62.0 \pm 14.7$ | $64.4 \pm 12.8$ | necklace | $51.3 \pm 22.5$ | $65.6 \pm 12.4$ |
| skirt | $59.4 \pm 10.4$ | $55.7 \pm 14.0$ | bracelet | $49.5 \pm 19.8$ | $50.2 \pm 22.8$ |

Table 4: Recall for selected garments on Fashionista.

(substituted with 'background'), both accuracy and mAGR values increase significantly (table 2) indicating that the algorithm had modest success in modelling the 'occluding object'. Figure 3 qualitatively compares our method with [16].

**Fashionista:** Yamaguchi *et al*. [16] have defined a protocol while evaluating the algorithm. The dataset has 56 labels and typically only a few of them are present in any given image. For each image, the identity of these labels are made available to the algorithm, without of course divulging the location of these labels. The algorithm is then expected to predict the location for each label in the image. As seen in the table 3, our method is marginally higher in mAGR and marginally lower in accuracy than the Yamaguchi *et al*. [16]. Table 4 shows the recall for several cloth types. Clearly our method is on par with [16] in most of the labels. Figure 4 qualitatively compares our method with [16].

**Co-occurrences:** Using the labelling obtained on Fashionista dataset reported above, cloth and color co-occurrences are computed as described in section 4. The minimum support threshold is set to 4%. The top 5 cloth co-occurrences for Fashionista dataset are skirt-top (6.3), shorts-top (5.7), blouse-skirt (4.5), tights-dress (4.4) and cardigan-dress (4.1). Similarly the top 5 cloth co-occurrences for Fashionista dataset are upper-blue:lower-blue (17.0), upper-red:lower-red(13.4), upper-red:lower-blue (12.0), upper-blue:lower-red (7.0) and upper-white:lower-blue (6.1). Figure 5 displays results of both cloth and color co-occurrences.

# 6 Conclusion

Understanding human clothing patterns and appearance has important consequences for human pose estimation, action recognition, surveillance, search and retrieval. Parsing for clothes in unconstrained settings is an important step towards better understanding of images. This paper proposes a method to segment clothes in images with no assumption on pose of the person, viewing angle, occlusion or clutter. An innovative method to model the

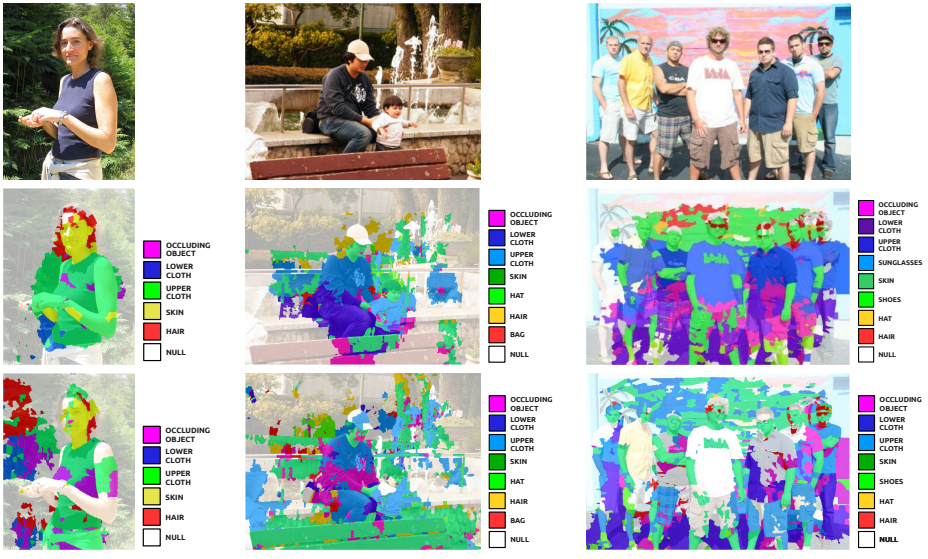


Figure 3: H3D results comparison: For the input images from the H3D dataset (row 1), results from our method (row 2) and Yamaguchi *et al*. [16] (row 3) are displayed. Clearly our method has superior segmentation than [16].

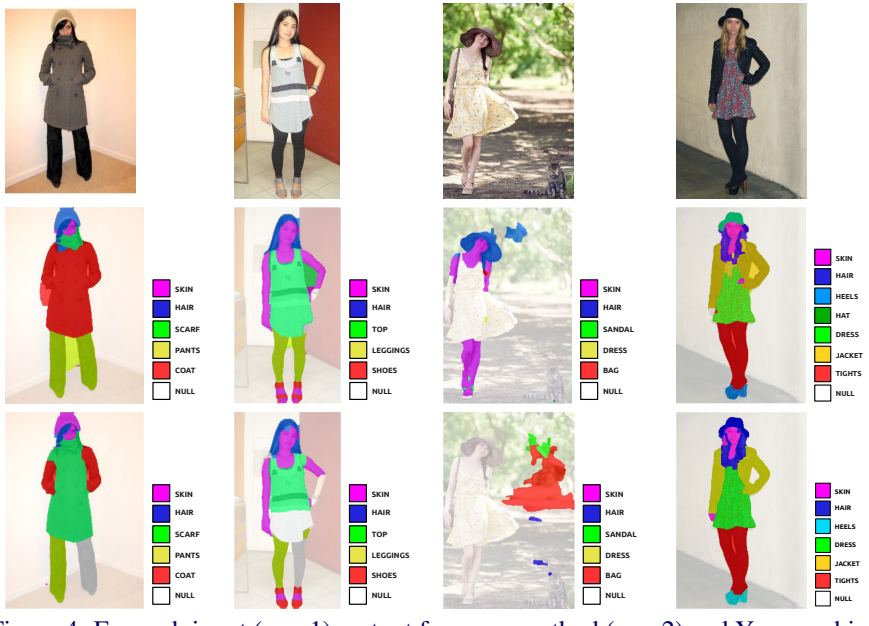


Figure 4: For each input (row 1), output from our method (row 2) and Yamaguchi *et al*. [16] (row 3) is displayed.

cloth-joint co-occurrence has been described which is invariant to the above challenges. The efficacy of the algorithm is demonstrated on challenging datasets and is shown to outperform the recent attempt [16].

(a) (b)

Figure 5: (a) Cloth co-occurrence: The first four columns correspond to Shorts-Top, Dress-Cardigan, Skirt-Top and Dress-Tights co-occurrences respectively. (b) Color co-occurrences of "upperbody cloth-lowerbody cloth" combination: The next four columns correspond to white-blue, blue-red, blue-blue and red-blue co-occurrences respectively.

# References

[1] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. *CVPR*, 2009.

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *CVPR*, 2009.

[4] Lubomir Bourdev, Subhransu Maji, Thomas Brox, and Jitendra Malik. Detecting people using mutually consistent poselet activations. In *European Conference on Computer Vision (ECCV)*, 2010. URL http://www.eecs.berkeley.edu/~lbourdev/poselets.

[5] George A. Cushen and Mark S. Nixon. Real-time semantic clothing segmentation. In *ISVC (1)*, 2012.

[6] Marcin Eichner and Vittorio Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009.

[7] Andrew C. Gallagher and Tsuhan Chen. Clothing cosegmentation for recognizing people. In *CVPR*, 2008.

[8] Basela Hasan and David Hogg. Segmentation using deformable spatial priors with application to clothing. In *BMVC*, 2010.

[9] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012.

[10] Joris M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:

2169–2173, August 2010. URL http://www.jmlr.org/papers/volume11/mooij10a/mooij10a.pdf.

[11] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. *ECCV*, 2010.

[12] Carl Scheffler and Jean-Marc Odobez. Joint adaptive colour modelling and skin, hair and clothes segmentation using coherent probabilistic index maps. In *BMVC*, 2011.

[13] Joseph Tighe and Svetlana Lazebnik. Superparsing - scalable nonparametric image parsing with superpixels. *IJCV*, 101(2):329–349, 2013.

[14] Nan Wang and Haizhou Ai. Who blocks who: Simultaneous clothing segmentation for grouping images. In *ICCV*, 2011.

[15] Michael Weber, Martin Bäuml, and Rainer Stiefelhagen. Part-based clothing segmentation for person retrieval. In *AVSS*, 2011.

[16] Kota Yamaguchi, M. Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.

[17] Ming Yang and Kai Yu. Real-time clothing recognition in surveillance videos. In *ICIP*, pages 2937–2940, 2011.

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