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

Cloth parsing involves locating and describing all the clothes (e.g., T-shirt, shorts) and accessories (e.g, bag) that the person is wearing. The main challenges in solving this include the large variety of clothing patterns that have been developed across the globe by different cultures. Occlusions from other humans or objects, viewing angle and heavy clutter in the background further complicates the problem.

In the recent past, Yamaguchi *et al*. [4] have proposed a method to parse clothes for fashion photographs where the image settings are simple with no clutter or occlusion. In our work, we aim to segment clothes in unconstrained settings by modelling the cloth to its body part vicinity in a CRF framework. Poselets [2] are adapted to obtain body part locations, as alternatives like human pose estimation algorithms frequently fail and give wrong pose estimates under occlusions and clutter.

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 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 probability function is given by,

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

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.

The unary potential function Φ 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 pairwise potentials, we use the definitions from [4]. Pairwise potential, defined between two neighboring super-pixels, model the interaction between them. The pair-wise potential is defined in equation 3 as sum of two functions (called factors) $\Psi(l_i,l_j)$ and $\Psi(l_i,l_j|Z,I)$. The pairwise potential function Ψ_1 models the likelihood of two labels l_i, l_j being adjacent to each other and Ψ_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 Ψ_1 is simply a log empirical distribution and Ψ_2 is 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,

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

Given images and cloth labels which include background, Logistic regression is used for $\Phi(l_i|Z,I)$ and $\Psi_2(l_i = l_j|\psi(s_i,s_j,Z))$. Given a new image, the super-pixels, poselets and feature vector ϕ 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 [3] package. The parameters λ_1, λ_2 in the equation 3 are found by cross validation. Our experiments on the complex H3D dataset [2] indicates that the proposed algorithm significantly outperformed the previous work [4] while on a relatively simple Fashionista dataset [4] it is on par.

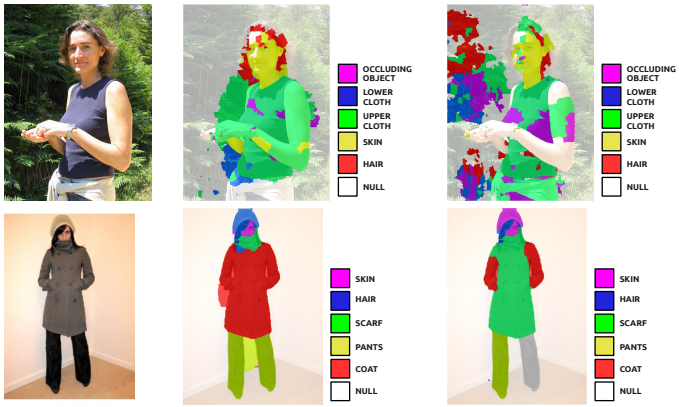

Figure 1: For the input image in column 1, our results are displayed in column2 and results of [4] are displayed in column 3.

Using the labelling obtained from the above method, interesting cloth and color co-occurrences can be mined using apriori algorithm [1].

Figure 2: **Cloth co-occurrences (Row 1):** The first three images display Cardigan-Dress co-occurrence and the next three images display Top-Skirt co-occurrence. **Color co-occurrence (Row 2):** The first three images display blue-blue co-occurrence and the next three images display white-blue co-occurrence.

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

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

[3] 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`.

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