Who Blocks Who: Simultaneous Clothing Segmentation for Grouping Images

Nan Wang and Haizhou Ai
Computer Science and Technology Department
Tsinghua University, Beijing, China
ahz@mail.tsinghua.edu.cn

Abstract

Clothing is one of the most informative cues of human appearance. In this paper, we propose a novel multi-person clothing segmentation algorithm for highly occluded images. The key idea is combining blocking models to address the person-wise occlusions. In contrary to the traditional layered model that tries to solve the full layer ranking problem, the proposed blocking model partitions the problem into a series of pair-wise ones and then determines the local blocking relationship based on individual and contextual information. Thus, it is capable of dealing with cases with a large number of people. Additionally, we propose a layout model formulated as Markov Network which incorporates the blocking relationship to pursue an approximately optimal clothing layout for group people. Experiments demonstrated on a group images dataset show the effectiveness of our algorithm.

1. Introduction

Clothing is one of the most informative cues of human appearance. It has been proven that clothing segmentation will benefit human detection [18] and recognition [9], pose estimation [16] and human sketches [5].

This paper addresses the problem of multi-person clothing segmentation with high occlusions, since numerous group pictures are captured in family gatherings and social activities. To analysis these images, inter-person occlusions must be taken into account.

Existing clothing segmentation algorithms [9][11] formulate the problem in the MRF framework with different clothing modeling approaches. In [11], clothing region (foreground) is initialized by torso detection based on dominant colors determination, while in [9], clothing mask is learned by mutual information analysis based on human identities. In addition, Gallagher et al. [9] show that clothing segmentation and human recognition can benefit each other. However, neither of the algorithms consider person-wise occlusions explicitly. Note that segmentation results in [9] seem to show its ability to address occlusions, but it is achieved by co-segmenting multiple images of the same person with the same clothes.

There have been a number of approaches to segment objects with partial occlusion. Some innovative papers use part-based models. Winn et al. [21] and Hoiem et al.[10] handle the object occlusions by using asymmetric pair-wise potentials, which pursue consistent layout output. Kumar et al. [15] allow for self occlusion by building a Layered Pictorial Structure for specific class. Clothing shapes differ from each other dramatically due to severe occlusion (Fig. 1). It will be more appropriate to choose global masks for each person, since local parts are difficult to learn using such inconsistent data.

A similar work to ours is the layered model introduced by Yang et al. [22], where occlusions are handled by object depth estimation. The layer order is searched in a brute force way based on detector confidence and bottom edge
locations. Group images can contain several dozens of people, which makes the brute force search intractable sometimes. Moreover, it is not necessary to determine the occlusion relationship for all pairs of persons.

Our work is also inspired by the work of Eichner et al. [7], where inter-person occlusions are modeled based on relative face location and image border to facilitate multi-person pose estimation. However, relative face location can hardly help the occlusion determination for people abreast. We argue that clothing self-similarity is more helpful in this case and integrate the two cues in the form of Random Forest.

In this paper, we propose to estimate the clothing shapes for group people with a global view of the scene. We treat each person as a node and model their interaction (blocking relationship) as edges. The reason why we use such a person-graph (Fig. 1) is that the variation of clothing shapes in group images can be distinguished by their individual and context information (Note that in this scenario, human poses do not change a lot, so clothing shapes are more affected by inter-person occlusions). Pixel-level segmentation can benefit from the top-down guide.

The key aspect of our method is the use of Blocking Model for occlusion reasoning. As mentioned before, it is intractable to search for the full layer order in a brute force way for group images. However, it is not necessary to recover the full order, since local blocking relationship is informative enough for clothing shape decision. Thus, in contrary to the traditional layered model used in [22], we partition the full layer ranking problem into a series of pair-wise ones. A Blocking Model combining individual and context information is built to infer the pair-wise layer order.

Taken the inferred blocking results as prior information, we propose a Layout Model formulated as Markov Network to pursue the approximately optimal clothing layout for group people. Therefore, the ambiguities in overlapping region are reduced to guide the bottom-up segmentation algorithm more accurately.

Our main contributions are: (1) an effective Blocking Model to estimate the inter-person occlusion, which integrates individual and context information and is capable of scaling to large number of people; (2) a Layout Model to generate approximately optimal clothing layout for group people conditioned on the inferred blocking relationship.

The rest of the paper is organized as follows: we describe our probabilistic model in Sec. 2 including the Layout Model and the Blocking Model, and introduce the detailed learning and inference algorithm in Sec. 3 and Sec. 4 respectively. Experiments are demonstrated in Sec. 5. Finally, we come to conclusion in Sec. 6.

2. Multi-Person Clothing Model

Inter-person occlusions in group images are usually severe and dense. An extreme case is shown in Fig. 1, where some people’s bodies are completely blocked. In such scenarios, the largely-blocked people can only located through face detection (rectangles in Fig. 1) rather than human detection.

Given an image and its $N$ faces being detected, our algorithm is aimed to infer the appropriate clothing shape for each person considering the occlusion between neighbors. The whole algorithm is illustrated in Fig. 2. Let $x_i$ denote the features extracted for each person, including face location and size, superpixel image [6] and RGB features in clothing region. Supposing $M$ candidate clothing shapes are generated for each person (as described in Sec 3, these candidate shapes are generated by Random Forest), shape label $y_i \in \{1, \cdots, M\}$ represents the clothing shape selected for each person.

![Graphical diagram of our algorithm. Input data are the image [8] and face detection results [12]. Our algorithm outputs not only the multi-person clothing segmentation results but also the blocking relationship between neighbors.](image-url)
The person pairs possibly occluded by each other are denoted as the edge set $\mathcal{E}$. For any pair of persons, binary blocking indicator $b_{ij}$ denotes the blocking relationship. When person $i$ blocks person $j$, $b_{ij} = 1$; otherwise, $b_{ij} = 0$. And $b_{ij} = 1 - b_{ji}$.

Let $x = \{x_1, \cdots, x_N\}$, $y = \{y_1, \cdots, y_N\}$ and $b = \{b_{ij} | (i, j) \in \mathcal{E}\}$. The multi-person clothing model maximizes the conditional probability $P(y, b|x)$ for clothing shape $y$ and blocking relationship $b$. The probability can be factored based on Bayes’ theorem:

$$P(y, b|x) = P(y|x, b) P(b|x)$$

$P(y|x, b)$ is the probability of each clothing layout $y$ given the blocking information $b$. We call it Layout Model. $P(b|x)$ is the probability of blocking relationships for current image. We call it Blocking Model.

### 2.1. Layout Model

We assume that the person-graph is a Markov Network, so the Layout Model can be factored as:

$$P(y|x, b) = \frac{1}{Z} \prod_i \varphi_i(y_i|x_i)^{\beta_u} \prod_{(i,j) \in \mathcal{E}} \varphi_{ij}(y_i, y_j|x_i, x_j, b_{ij})^{\beta_{ij}}$$

where $Z$ is the partition function to make sure that $P(y|x, b)$ is a distribution. $\mathcal{E}$ denotes the graph structure, which will be decided in runtime based on face locations and sizes.

The unary potentials $\varphi_i$ measure the consistency between the clothing shape and the actual edges. We segment the image into superpixels [6]. The region boundaries are used as the actual edges of the image. We perform a naive inference in pixel and superpixel level respectively based on the clothing shape. Suppose the inferred foregrounds in pixel and superpixel level are $L_{px}$ and $L_{rg}$ respectively, then $\varphi_i$ is defined as their consistency:

$$\varphi_i = \frac{|L_{px} \cap L_{rg}|}{|L_{px} \cup L_{rg}|}$$

A uniformed unary parameter $\beta_u$ (Eq. 2) is assumed for all unary potentials.

The pairwise potentials $\varphi_{ij}$ penalize the conflict foreground part of the neighbor shapes and encourage larger size of the unoccluded one. Denoting the overlapping region as $A$, $\varphi_{ij}$ is calculated as:

$$\varphi_{ij}(y_i, y_j|b_{ij}) = \exp(-\varphi_{ij}^b - \varphi_{ij}^c)$$

where

$$\varphi_{ij}^c = \sum_k y_{ij}^k y_j^k / |A|$$

and

$$\varphi_{ij}^b = \frac{\sum_k y_{ij}^k b_{ij} + \sum_k y_j^k (1 - b_{ij})}{\sum_k y_{ij}^k + \sum_k y_j^k}$$

$\beta_{ij}^b$ is the weight parameter of pairwise potentials. In our method, $\beta_{ij}^u$ represent the confidence of the blocking relationship. So we set $\beta_{ij}^u = P(b_{ij})$, where $P(b_{ij})$ is the blocking probability calculated as Eq. 10. $y_i^k$ is the binary indicator of shape of pixel in $A$, the same to $y_j^k$; $|A|$ is the size of the overlapping region. In Eq. 5, the numerator accumulates the number of pixels where occluded and unoccluded shapes are foreground simultaneously. This conflict of foreground brings the ambiguity for bottom-up segmentation. In Eq. 6, we encourage a bigger foreground size of the unblocked person in the overlapping region by calculating their foreground size proportions. For both of the potentials, smaller values are preferred.

### 2.2. Blocking Model

Heuristically, we can say the person with a lower face location is more likely to block that with a higher one in a group image. However, estimating the occlusion relationship by relative face location is not as reliable as human body locations [22]. For example, the 0-th and 2-th face locations in Fig. 1 are lower than their neighbors, but they actually stand behind them. It is not a rare case, since in group images, people prefer to stand in rows. A helpful observation is that unoccluded clothing usually is symmetric, or contains more repetitive patterns. The two cues are used in our model to estimate the blocking probability.

We assume that the blocking model for all pairs of neighbors are independent to each other. so $P(b|x)$ can be factored as:

$$P(b|x) = \prod_{(i,j) \in \mathcal{E}} P(b_{ij}|x_i, x_j)$$

The blocking probability between pairs of persons takes the form of random forests as shown in Fig. 3.

### 3. Model Learning

We learn the model using a supervised algorithm based on a manually labeled data set. Different persons are labeled respectively. The blocking relationship is generated from the segmentation labeling results. For nearby persons whose clothing regions (estimated by face location and size) overlap the other, we accumulate the foreground size of them in the overlapping region and determine the blocking relationship $b_{ij}$ and confidence $P(b_{ij})$ based on the foreground proportion:
Given an image and its face detection results, we are aiming at inferring the clothing layout and also the blocking relationship for group people.
First, as suggested by Kumar et al. [15], when we use the top-down model as latent variables to guide the bottom-up image segmentation, the expectation of the log likelihood of an MRF can be efficiently optimized by a single graph cut optimization. Importantly, they argue that multiple samples are necessary for some difficult cases in which RGB distribution of background is similar to that of the object.

The other is from an insight of our model. In Sec 3.1, a clothing shape random forest is learned for each person individually. Random forest increases its generalization accuracy nearly monotonic with respect to the number of decision tree [4]. The final voting procedure of outputs will make the model more robust to noise. However, the MAP inference selecting only one result from all tree outputs will decrease the generalization ability. From this point of view, sampling from the Layout Model, where interaction of the neighbors is taken into account, is a novel effective voting method to improve the individual random forest performance.

For convenience of sampling, we use a sum-product algorithm for MAP inference [14] and then sample multisolutions around the local (global for tree-graph) minimum by Gibbs sampling. Here we approximate the posterior distribution of layout solution as [15]:

$$p(y) = \frac{\prod_{i,j \in E} B_{ij}(y_i, y_j)}{\prod_i q_i^{q_i}}$$

where $B_i$ and $B_{ij}$ are the unary and pairwise beliefs calculated by the sum-product algorithm. $q_i$ is the neighbor number of node $i$.

5. Experiments

We evaluate our method on a public available data set of group images [8]. All the images with high occlusions between neighbors are downloaded from Flickr by different key words: wedding, family and group images. We manually label 281 images with 1051 persons. We partition the ground truth dataset into two halves randomly and train our model using one half and test it with the other. In our experiments, clothing is labeled as any covering of the torso, including naked shoulder and arms.

Implementation: We detect face location and scale using face detector [12] and normalize the images to make each face $24 \times 24$. $\beta^u$, the weight of unary potentials in Eq. 2 is set as 0.4 empirically by cross-validation.

Random Forests for clothing shape and blocking relationship consist of 50 decision trees. For each node of them, we randomly select $\sqrt{N_f}$ [2] weak clothing self-similarity features (relative face location features are used for all nodes), where $N_f$ is the weak feature number (around 300,000). The size of rectangles used in clothing self-similarity features is $4 \times 4$, and they are densely sampled

4.1. Data Preparing

Based on face locations, we estimate the possible clothing regions as rectangles. The edges $E$ for Layout Model are determined for the neighbor rectangles with large enough overlapping regions (10% of the minimum rectangle).

For each person, we obtain the candidate clothing shapes using our clothing shape random forest learned in Sec. 3.1. For each pair of neighbors, we obtain the blocking distribution using Eq. 10.

4.2. Blocking Relationship

The optimal blocking relationship $\hat{b}$ can be obtained by maximizing the individual local blocking distribution, i.e.

$$\hat{b} = \prod_{(i,j) \in E} \arg \max_{b_{ij}} P(b_{ij}|x_i, x_j)$$

Yang et al. [22] search the optimal ordering from all possible permutation. In our method, we just use the MAP solution, since we observe that the unsure blocking relationship usually occurs when neighbors stand close but do not block each other (as shown in Fig. 4). It means that "who blocks who" is not critical in these cases. We incorporate the confidence $P(b_{ij}|x_i, x_j)$ to Layout Model in Eq. 4 by the pairwise weight, which decreases the effect of neighbor constraints when the relationship is not so confident.

4.3. Sampling Layout

Given blocking configuration $t$, we can approximate the MAP solution of Eq. 2 by Loopy Belief Propagation. However, in our method, we propose to sample the layout instead of a single MAP solution. The two reasons are as follows:

Figure 4. To get a detailed view of the performance of our blocking model, we partition the testing samples into different groups based on the proportion of foreground sizes of neighbors in the overlapping region ($\min(A, B) / (A + B)$, referring to Eq. 6). The smaller the proportion is, the more likely they block each other. Based on the plot, we find that our algorithm performs well for severe blocking cases (88%). For the inaccurate proportion, we observe that the blocking relationship is not so obvious (see the representative images in the figure), so we depress the inaccurate effect by incorporating the confidence into the pairwise potential functions (in Eq. 4), since it (red curve) is highly correlated with the accuracy (blue curve).
Algorithm 1: Layout optimization for group images

**Input:** Image data and face detection results  
**Output:** Probabilistic clothing mask  

**[Data Preparing]**

1. Determine the inter-person relationship based on face locations and construct the graph;  
2. Generate shape candidates \( \{q_i\} \) for each person based on clothing model in Sec. 3.1;

**[Model Inference]**

1. Estimate the blocking relationship \( \{b_{ij}\} \) and its confidence \( \{P(b_{ij})\} \);  
2. Optimize Eq. 2 by Loopy Sum-Product Belief Propagation to get the MAP inference result, the unary beliefs \( \{B_i\} \) and pairwise beliefs \( \{B_{ij}\} \);  
3. Generate multiple solutions around the MAP one by Gibbs sampling (Eq. 12);  
4. Average the multiple clothing shapes for each person to generate the probabilistic clothing mask;

with step 2. We calculate the upper and lower bound of feature values for each node and partition the interval into 10 pieces and then select the optimal threshold from them. No pruning is made for decision trees.

A sketch of our layout optimization algorithm is shown in Algorithm 1. Only original image data and face detection results are required. The algorithm outputs a probabilistic mask for each person providing top-down guide for further pixelwise segmentation.

To benefit the bottom-up segmentation, we incorporate the top-down cues as a unary term in a CRF based framework [20]. Given the clothing shapes for each person by sampling Eq. 12, we obtain the multi-person clothing mask by a Bayesian fusion as:

\[
p(l_i = 0) = \prod_{m=1}^{M} p(l_i = 0 | y_m) \tag{13}
\]

\[
p(l_i = m) = \frac{p(l_i = 1 | y_m)}{\sum_k p(l_i = 1 | y_k)} \cdot (1 - p(l_i = 0)) \tag{14}
\]

where \( l_i \in \{0, 1, \ldots, M\} \) is the label of pixel \( i \), in which 0 is the background and \( \{1, \ldots, M\} \) are different persons. Intuitively, \( l_i = 0 \) if and only if in each person’s clothing mask, pixel \( i \) is background.

Clothing and background color models (histograms in RGB space) are estimated based on the clothing masks. Shape and color cues are integrated with superpixel-level constraints [6]. The final segmentation result is obtained by optimizing a Dual-Level Conditional Random Field [20].

**Blocking Modeling:** The blocking performance is evaluated for the ones whose confidence \( P(b_{ij}^p) > 0.7 \), since the blocking relationship is not so critical for the others, as show in Fig. 4.

Experiments are performed with different configurations: only face information, only clothing information and both of them (full model). Based on the curve in Fig. 5, we find that face information achieves baseline accuracy, since it portrays the basic structure of group people. However, it cannot handle the occlusion between several people abreast, which limits its capability for blocking relationship prediction. Moreover, its performance will not increase with respect to the number of trees, since actually there are not so many available features for our random selection.

Both clothing information and our full model can achieve better accuracies with large enough amount of trees. However, the experiments show that to achieve the same performance, the full model needs less trees. In addition, as shown in the right curves of Fig. 5, the size of random forest based on both kinds of information is much smaller than that based on only clothing information. So the conclusion is that the clothing information can capture the blocking relationship between persons, while face information can enhance it by reducing the required number of trees and nodes. As a comparison, we implement the method in [7]. However, its MAP result only achieves 56.2% accuracy. The reason may be that in our dataset, people stand closer which reduce the power of *Face Relative Size* and moreover no self-similarity features are explored.

**Layout Modeling:** We investigate the performance of Layout Model by segmentation accuracy based on the manually labeled dataset.

Tab. 1 reports pixelwise occlusion error (the proportion of pixels assigned to wrong persons) and segmentation accuracy. We apply three different methods to segment the clothing, which gives a detailed view of the performance of our model. As a baseline method, we average over all clothing shapes to get a global mask, and just scale the probabilistic mask for each person with respect to the face size. In this case, no blocking information is used. The perfor-
Table 1. Segmentation performance comparison. Single mask: averaging over all clothing shapes, which uses none occlusion cues. CF: only use clothing shape forest, which addresses blocking information implicitly. CF + BF: Clothing Forest + Blocking Forest (integrated by Layout Model), which addresses blocking relationship explicitly.

<table>
<thead>
<tr>
<th>Method</th>
<th>Occlusion Error</th>
<th>Pixel Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Mask</td>
<td>6.8%</td>
<td>88.2%</td>
</tr>
<tr>
<td>CF</td>
<td>5.9%</td>
<td>90.7%</td>
</tr>
<tr>
<td>CF + BF</td>
<td>4.6%</td>
<td>92.8%</td>
</tr>
</tbody>
</table>

Performance can be improved by using our Clothing Forest model, which is constructed based on clothing self-similarity features. As stated before, unoccluded clothing usually contains more repeatable patterns, so in fact Clothing Forest incorporates blocking information implicitly. Occlusion between persons can be handled moderately. An example is shown in Fig. 6, which is marked with an orange single-boundary rectangle. However, when assigning a putative clothing shape to each person, Clothing Forest still works in an independent way. Our Layout Model enhances this procedure by optimizing the layout in a global view of the entire people group. Blocking information is captured by Blocking Forest and embedded as pairwise constraints.

![Figure 6](image_url)

Figure 6. Comparative segmentation results. (a) input image and the face detections (b) results using single clothing mask (c) results using clothing shape random forest directly (d) results by incorporating the blocking relationship.

Some of the comparison segmentation results produced by the three methods are shown in Fig. 6. Due to the help of blocking information, our segmentation algorithm can produce much better object-level boundaries even when the clothing of different persons is very similar (the occlusion parts are marked by double-boundary rectangles). Also, we report the pixelwise segmentation accuracy under each segmentation result. We can see that although the quantitative improvement in reported accuracy seems to be marginal, for segmentation problem, a small increment in pixelwise accuracy can produce remarkable qualitative segmentation improvement, which is also stated in [19][13]. We also compare to GrabCut [17] (implemented by OpenCV [2]) in Fig. 7. The initial rectangle is generated based on global clothing template. Without blocking information, GrabCut might fail for blocked people.

We show additional segmentation results from the data set [8] in Fig. 8. In addition, as stated before, we partition the full layer sorting problem into pairwise ones, so our algorithm is capable of dealing with cases with a large number of persons. Fig. 8 also shows such examples. Our algorithm produces nice segmentation results even in such a crowded scene. Note that inaccurate segmentations occur in the arm parts, since the color of the skin is different from that of clothing while we do not filter that out especially.

6. Conclusion

In this paper, we have proposed a novel approach for multi-person clothing segmentation problem. Blocking information is incorporated in a principle way to benefit a clothing shape sampling procedure. Experiments on severe crowded group image with arbitrary number of people show that our approach outperforms conventional method and yield expressive segmentation results.

Currently, the model assumes that there are not serious pose variations. Thus, there is clearly room for improvement, particularly in combining with human pose estimation (e.g. multiple-person pose estimation [7], simultaneous pose estimation and human segmentation [3][1]).

7. Acknowledgements

This work is supported in part by National Science Foundation of China under grant No.61075026, National Basic Research Program of China under Grant No.2011CB302203, and it is also supported by a grant from HP Corporation.
Figure 8. Some additional segmentations. For each pair of images, the first one is the input image with detected face, and the other is the segmentation result produced by our algorithm. The last two rows show expressive segmentation results for crowded scene.

References