Supplemental Material: Discovering Groups of People in Images

Wongun Choi[1], Yu-Wei Chao[2], Caroline Pantofaru[3] and Silvio Savarese[4]

1. NEC Laboratories 2. University of Michigan, Ann Arbor
3. Google, Inc 4. Stanford University

1 Qualitative Examples

In Fig. 1 and 2, we show additional qualitative examples obtained using our model with poselet [1] and ground truth (GT) detections, respectively. We show the image configuration of groups on the left and corresponding 3D configuration on the right. Different colors and different line types (solid or dashed) represent different groups, the type of each structured group is overlayed on the bottom-left of one participant. In 3D visualization, squares represent standing people, circles represent people sitting on an object, and triangles represent people sitting on the ground. The view point of each individual is shown with a line. The gray triangle is the camera position. The poses are obtained by using the individual pose classification output for visualization purposes.

The figures show that our algorithm is capable of correctly associating individuals into multiple different groups while estimating the type of each group. Notice that our algorithm can successfully segment different instances of the same group type that appear in proximity. A distance-based clustering method would not be able to differentiate them. The last figure shows a typical failure case due to only reasoning about people while ignoring objects (such as the tables). Also, we notice that our algorithm can associate individuals into correct groups even in highly complicated scene when GT detections are available.

2 Dataset Statistics

In this section, we analyze the statistics of the newly proposed Structured Group Dataset. The dataset is composed of 588 images with 5,415 human annotations and 1,719 groups (excluding outliers). We mirror all the images to get 1,176 images with 10,830 humans and 3,438 groups. The groups are categorized into 7 different types of structured groups; 1) *queuing* (Q), 2) *standing facing each other* (SF), 3) *sitting facing-each-other* (OF), 4) *sitting on the ground facing-each-other* (GF), 5) *standing side by side* (SS), 6) *sitting side by side* (OS), and 7) *sitting on the ground side by side* (GS). We show the summary statistics of our dataset in Tab. 1 and Fig. 3. The statistics show that the groups in the SGD dataset have a high amount of variation in the number of group participants.

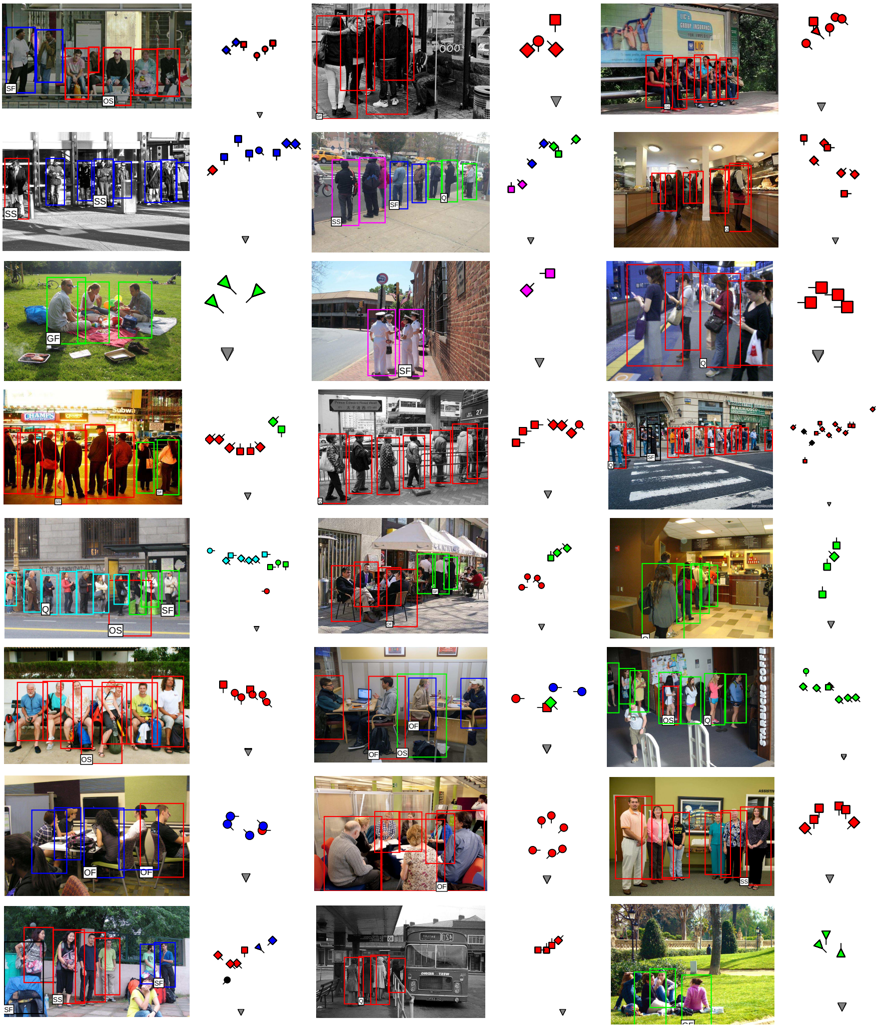

Fig. 1: Qualitative examples of the results obtained using our full model with Poselet detections [1]. See text for the details.

3 3D Estimation from Single Image

3.1 Model

Given an image I, we estimate the camera parameter Θ and people in 3D Z using a technique similar to [3]. Denote $\Theta = \{f, \phi, h_c\}$, where f is the camera focal length, ϕ is the pitch angle, and h_c is the camera height (we assume zero yaw and roll angles). We model the full body and torso of each person as pose-dependent cuboids in the 3D space. Assuming we have N detected people, denote $Z = \{z_1, \cdots, z_N\}$, where each person z_i

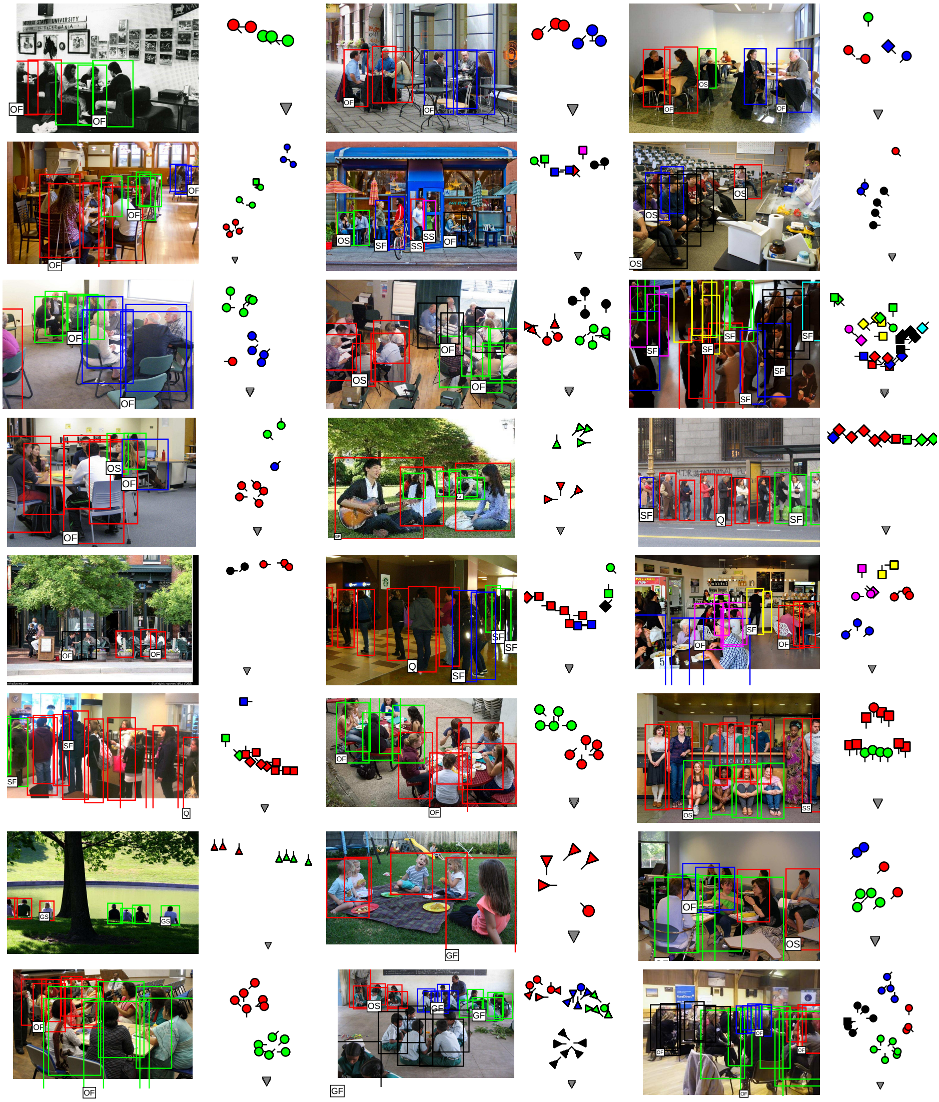

Fig. 2: Qualitative examples of the results obtained using our full model with GT detections. See text for the details.

is represented by the 3D location of the cuboid bottom $c_i \in \mathbb{R}^3$, pose $b_i \in \{1, 2, 3, 4\}$ (standing, sitting on an object, sitting on the ground, and false positives), and height $h_{z_i} \in \mathbb{R}$. The inputs of our system are 1) N detected human returned by the Poselet detector [1] (characterized by a full body bounding box and a torso bounding box, see Fig. 5), 2) the geometric context feature [2] extracted from I, and 3) prior distributions on the cuboid sizes for different poses. Our system outputs Θ and Z by solving the following energy maximization problem:

$$E(\Theta, Z, I) = \omega_{\Theta I}\Psi(\Theta, I) + \omega_{\Theta Z}\Psi(\Theta, Z) + \omega_{ZI}\Psi(Z, I) + \omega_{\Theta}\Psi(\Theta), \tag{1}$$

where $\Psi(\Theta, I)$ captures the compatibility between the camera parameter Θ and the image feature. $\Psi(\Theta, Z)$ captures how well the humans in configuration Z fit into the scene given the camera parameter Θ. $\Psi(Z, I)$ captures how likely the human configuration Z is given the observed image. $\Psi(\Theta)$ accounts for the prior on Θ. $\omega_{\Theta I}$. $\omega_{\Theta Z}$, ω_{ZI}, and ω_{Θ} are the model weight parameters.

Image-Camera Compatibility $\Psi(\Theta, I)$: This potential measures the compatibility between the geometric context feature [2] extracted from I and the camera parameter Θ. Let (u, v) denote the indices of x and y coordinates on the I. Given f, ϕ, and the camera principle point (u_c, v_c), we can compute the horizon line position v_0 in I by

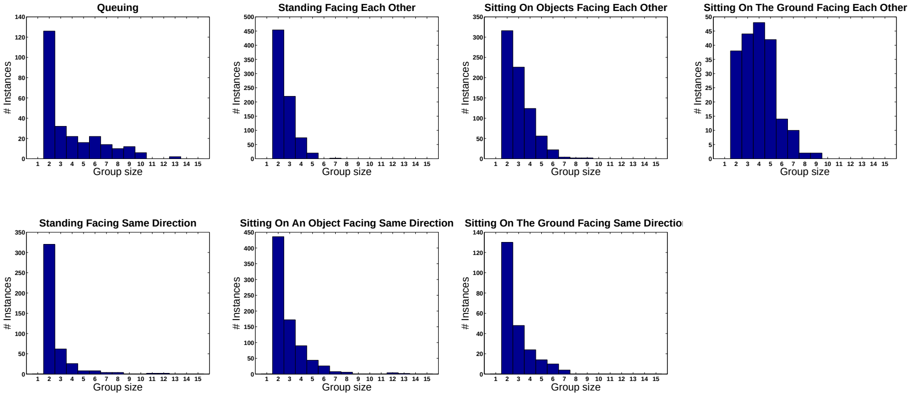

Fig. 3: The distribution of the number of people in each group. Our dataset has a high amount of variation in group configuration.

Group Type	Q	SF	OF	GF	SS	OS	GS
# instances	262	770	752	200	436	788	230
Mean size	3.89	2.56	3.03	3.99	2.57	2.92	2.86
STD size	2.48	0.80	1.18	1.53	1.38	1.48	1.25

Table 1: Summary statistics of our dataset. We show the number of group instances, the mean number of group participants (Mean size), and the standard deviation of the number of group participants (STD size) per each group type. The groups in our dataset have high configuration variation. See Fig. 3 for the histograms.

$v_0 = v_c - f\tan(\phi)$ as shown in Fig. 4. The potential $\Psi(\Theta, I)$ is formulated as:

$$\Psi(\Theta, I) = \frac{1}{N_{pix}} \sum_{u} \left(\sum_{v<=v_0} p_{sky}(u,v) + \sum_{v>v_0} p_{sup}(u,v) \right), \tag{2}$$

where $p_{sky}(u,v)$ and $p_{sup}(u,v)$ are the probabilities of the pixel at (u,v) belongs to the geometric class *sky* and *support*, respectively, using [2]. N_{pix} is the total number of pixels in I.

Camera-Human Compatibility $\Psi(\Theta, Z)$: This potential measures the likelihood of the human configuration Z given the camera parameter Θ. Assuming z_is are independent, we have,

$$\Psi(\Theta, Z) = \frac{1}{N} \sum_{i=1}^{N} \Psi(\Theta, z_i), \tag{3}$$

where $\Psi(\Theta, z_i)$ measures how close the 3D height h_{z_i} is to the expected height of the full body cuboid given Θ and pose b_i. Assuming the cuboids and the ground plane have the same normal, we can obtain the person's 3D location c_i and height h_{z_i} by the following process: 1) get the person's depth by back-projecting the torso bounding box until it fits the height of the torso cuboid model, 2) get the bottom of the person by extending the torso until it touches the ground, and 3) get the top of the person until it intersects with the back-projecting ray of the top of full body bounding box. Fig. 4 illustrates this process. Note that we observe the bottoms of the detected full body bounding boxes are in general very noisy on the *Structured Group Dataset*, but the torso region and the top of the full body bounding boxes are mostly accurate, so we rely on these two features to obtain a robust estimation on human depth. Once we have h_{z_i}, $\Psi(\Theta, z_i)$ is formulated as

$$\Psi(\Theta, z_i) = \begin{cases} \ln \mathcal{N}(h_{z_i} - \mu_{b_i}, \sigma_{b_i}) & \text{if } b_i \in \{1,2,3\} \\ \ln \alpha & \text{if } b_i = 4, \end{cases} \tag{4}$$

where μ_{b_i} and σ_{b_i} characterize the distribution of the full body height for pose b_i, and α is a constant value used for false positives. In practice, we set μ_{b_1}, σ_{b_1}, μ_{b_2}, σ_{b_2}, μ_{b_3},

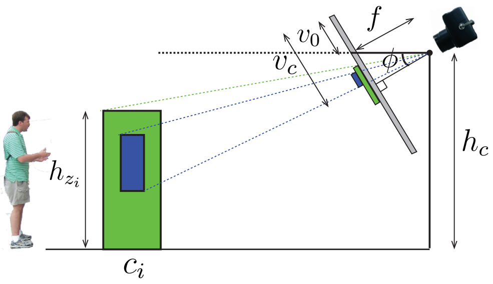

Fig. 4: Illustration of our 3D estimation method. The green cuboid represents the full body and the blue cuboid represents the torso.

Algorithm 1 RANSAC algorithm for solving Θ

```
while $count \leq maxiter$ do
    $Z_s \leftarrow$ Sample 3 people bounding boxes and their poses from the detection set
    solve $E_s = \arg\max_{\Theta} E(\Theta, Z_s, I)$
end while
return $\Theta_s$ corresponding to the highest $E_s$
```

σ_{b_3} and α to be 1.68, 0.10, 1.32, 0.10, 0.89, and 0.10 (all in meters). The 3D torso height is set to be 0.5 (meters) for $b = 1$ or 2, and 0.45 (meters) for $b = 3$.

Image-Human Compatibility $\Psi(Z, I)$: The compatibility between human Z and image I is defined by the pose classification confidence as,

$$\Psi(I, Z) = \frac{1}{N} \sum_{i=1}^{N} \ln p_{pose}(b_i), \tag{5}$$

where $p_{pose}(b_i)$ is the probability of ith detection having pose b_i returned by the trained pose classifier.

Camera Prior $\Psi(\Theta)$: We apply the following prior on the camera parameters f, ϕ, h_c.

$$\Psi(\Theta) = \ln \mathcal{N}(f - \mu_f, \sigma_f) + \ln \mathcal{N}(\phi - \mu_\phi, \sigma_\phi) + \ln \mathcal{N}(h_c - \mu_{h_c}, \sigma_{h_c}). \tag{6}$$

In practice, we set $\mu_f, \sigma_f, \mu_\phi, \sigma_\phi, \mu_{h_c}$, and σ_{h_c} to be 550 (pixels), 100 (pixels), 8 (degree), 8 (degree), 1.68 (meters), and 0.6 (meters), respectively.

3.2 Inference

We propose to solve problem 1 using a two-step optimization: 1) first solve Θ using the RANSAC algorithm, and 2) solve Z by maximizing E given the optimized Θ. In the first step, we iteratively sample three people and their poses and fit a best Θ by maximizing E. This can be solved with a simplex search method [6]. After we have generated enough samples, we obtain the best Θ associated to the highest score E. This is detailed in Alg. 1. Once we solve Θ, we can obtain Z (3D locations, heights, and poses of people) by finding the pose b_i for each person that maximizes E.

3.3 Example result

In Fig. 5, we present example results of 3D estimation. The first row shows the input image. The second row shows the our input detection obtained by Poselet [1]. The third row shows the horizon line and the true positives returned by our algorithm. The last row shows the results of our system in 3D, from a top-down view. These example results demonstrate that our 3D estimation method is capable of removing false positives and generating 3D maps of people which are robust to 2D bounding box noise.

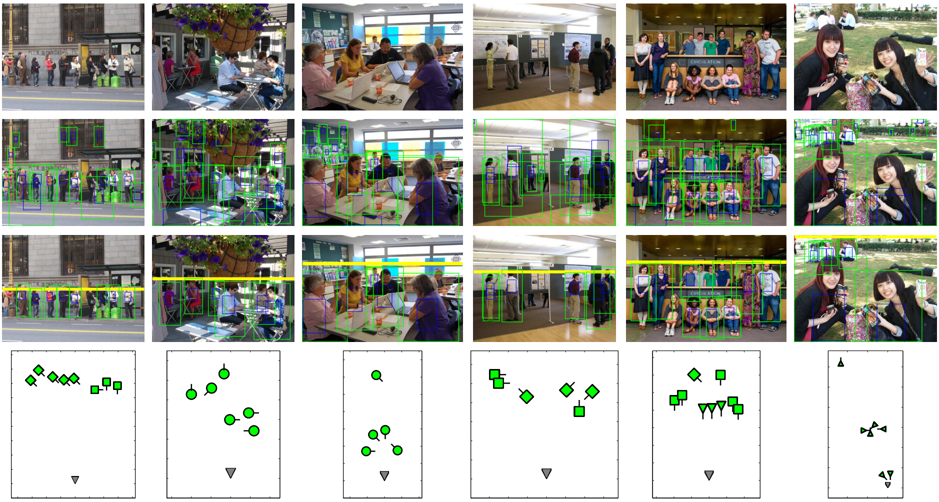

Fig. 5: Example results of our 3D estimation. The first row shows the input image. The second row shows the input Poselet detection [1]. The full body bounding boxes are colored green and the torso bounding boxes are colored blue. Note that the input detections are often noisy, e.g. bad localization and truncation by the image. The third row shows the horizontal line (yellow) and the true positives returned by our method. Our method is able to remove false positives. The last row shows the 3D map in a top-down view. Squares represent standing people, circles represent people sitting on an object, and triangles represent people sitting on the ground. The view point of each individual is shown with a line. The gray triangle is the camera position. This shows our method's ability to generate 3D estimates which are robust to noisy 2D bounding boxes.

4 Inference with Mean Field Message Passing

As described in Sec. 4 of the paper, we obtain our solution by optimizing over the following objective function iteratively:

$$\nabla\Psi(C_k;\hat{\mathbb{C}}_{k-1},\mathbb{X},\mathbb{Y}) = \Psi(\hat{\mathbb{C}}_{k-1}\oplus C_k,\mathbb{X},\mathbb{Y}) - \Psi(\hat{\mathbb{C}}_{k-1},\mathbb{X},\mathbb{Y}) \tag{7}$$

where $\mathbb{Y}$ are given and $\hat{\mathbb{C}}_{k-1}$ is given in the previous iteration. The new group $\hat{\mathbb{C}}_k$ is obtained by using the augmentation operator $\hat{\mathbb{C}}_{k-1}\oplus C_k$. We optimize Eq. 7 by applying a variational method on each group type $\hat{c}$. Fixing the group type $\hat{c}$, the optimization space can be represented by the membership vector $\hat{H}_k$. With a slight abuse of notation, we can reformulate the optimization problem with a fully connected conditional random field (CRF) as:

$$\nabla\Psi(\hat{H}_k) = \sum_i \psi_u(\hat{h}_i^k) + \sum_{i<j} \psi_p(\hat{h}_i^k,\hat{h}_j^k) \tag{8}$$

We describe the details of reformulation and MF-MP algorithm in following sections.

4.1 Objective Function of MF-MP objective

Given current $\hat{\mathbb{C}}_{k-1}$ and interaction variable $\mathbb{Y}$, we can rewrite the unary potential ψ_u and pairwise potential ψ_p following the Eq.7. As the first, we define the current group type assignment for each individual γ_i and the current group type assignment for each pair $\gamma_{i,j}$ that can be obtained from $\hat{\mathbb{C}}_{k-1}$, i.e. if an individual detection i is included in a group C_m, $\gamma_i = c_m$, otherwise $\gamma_i = B$ and if both of i and j are included in a group C_m, $\gamma_{i,j} = c_m$, otherwise $\gamma_{i,j} = B$. Given the group type assignment γ, we can write the unary potential ψ_u as follows:

$$\psi_u^{k-1}(\gamma_i) = I(\gamma_i, B)w_{xb}^\top \psi_{xb}(x_i) + (1 - I(\gamma_i, B))w_{xc}^\top \psi_{xc}(x_i, \gamma_i) \tag{9}$$

$$\psi_u(\hat{h}_i^k) = \hat{h}_i^k(w_{xc}^\top \psi_{xc}(x_i, c_k) - \psi_u^{k-1}(\gamma_i)) \tag{10}$$

where $\psi_u^{k-1}(\gamma_i)$ is the unary potential contribution from the previous group assignments and $I(\cdot,\cdot)$ is an indicator function. This unary potential measures the improvement in unary potential by assigning a new group type to individual detections. Similarly, we can write the pairwise potential ψ_p as follows.

$$\psi_p^{k-1}(\gamma_{i,j}) = I(\gamma_{i,j}, B)w_{yr}^\top \psi_{yr}(y_{i,j}) + (1 - I(\gamma_{i,j}, B))w_{yc}^\top \psi_{yc}(y_{i,j}, \gamma_{i,j}) \tag{11}$$

$$\psi_p(\hat{h}_i^k, \hat{h}_j^k) = \begin{cases} 0 & if\ \hat{h}_i^k = 0,\ \hat{h}_j^k = 0 \\ w_{yr}^\top \psi_{yr}(y_{i,j}) - \psi_p^{k-1}(\gamma_{i,j}) & if\ \hat{h}_i^k = 0,\ \hat{h}_j^k = 1 \\ w_{yr}^\top \psi_{yr}(y_{i,j}) - \psi_p^{k-1}(\gamma_{i,j}) & if\ \hat{h}_i^k = 1,\ \hat{h}_j^k = 0 \\ w_{yc}^\top \psi_{yc}(y_{i,j}, c_k) - \psi_p^{k-1}(\gamma_{i,j}) & if\ \hat{h}_i^k = 1,\ \hat{h}_j^k = 1 \end{cases} \tag{12}$$

where $\psi_p^{k-1}(\gamma_{i,j})$ is the pairwise potential contribution from the previous group assignments. Notice that selecting only one of the pairs in the new group assignment, force the interaction be repulsive (the second and third conditions in Eq. 12).

4.2 MF-MP algorithm

Given a fully connected CRF with unary and pairwise potentials, we can solve the problem using the Mean Field Message Passing Algorithm [5]. Define variational distribution Q as follows:

$$Q(H_k) = \prod_{h_i^k \in H_k} Q(h_i^k) \tag{13}$$

Then, we can find the solution H_k by maximizing over the variational distribution Q that minimize $KLD(P, Q)$ where $P(H_k) = \frac{1}{Z} exp(-\nabla \Psi(H_k))$. We can solve this problem by algorithm.2, where $scope(\psi)$ returns all the variable X that is an argument of ψ and $E_{Q_{\setminus X}}$ means expectation over the residual variational distribution $Q_{\setminus X} = \prod_{Y \in \mathbb{X} - X} Q(Y)$.

5 Structural SVM Training

We define the loss function $\delta(\mathbb{C}, \mathbb{C}_i)$ by accumulating individual group type association loss that is defined as follows:

$$\delta(\mathbb{C}, \mathbb{C}_i) = \sum_{n \in \mathbb{X}} (1 - I(\gamma^n, \gamma_i^n)) \tag{14}$$

Algorithm 2 Mean Field Message Passing Algorithm

```
$Q \leftarrow Q_0$
$\mathbb{X} = H_k$
$Unprocessed \leftarrow H_k$
while $Unprocessed \neq \emptyset$ do
    Choose $X \in \mathbb{X}$ from $Unprocessed$
    $Q_{old}(X) \leftarrow Q(X)$
    for $x \in Val(X)$ do
        $Q(x) \leftarrow exp\{\sum_{\psi : X \in scope(\psi)} E_{Q \setminus X}[log(\psi(x))]\}$
        Normalize $Q(X)$ to sum to one
    end for
    if $Q_{old}(X) \neq Q(X)$ then
        $Unprocessed \leftarrow Unprocessed \cup neighbor(X)$
    end if
    $Unprocessed \leftarrow Unprocessed - X$
end while
```

where γ^n is the group label induced from the group association $\mathbb{C}$ as described in Sec. 4.1 and γ_i^n is the group label induced from the ground truth group association. We penalize the configuration which associate individuals into wrong group categories. We also experimented with using a pairwise loss, but it did not improve the model learning significantly. Given the definition of the loss function, we optimize the model parameters using cutting plane algorithm introduced in [4].

References

1. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: International Conference on Computer Vision (ICCV) (2009), http://www.eecs.berkeley.edu/ lbourdev/poselets
2. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. IJCV (2007)
3. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. IJCV (2008)
4. Joachims, T., Finley, T., Yu, C.N.: Cutting-plane training of structural svms. Machine Learning (2009)
5. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
6. Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E.: Convergence properties of the nelder–mead simplex method in low dimensions. SIAM J. on Optimization 9(1), 112–147 (May 1998)