

Discovering Groups of People in Images

Wongun Choi¹, Yu-Wei Chao², Caroline Pantofaru³ and Silvio Savarese⁴

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

Abstract. Understanding group activities from images is an important yet challenging task. This is because there is an exponentially large number of semantic and geometrical relationships among individuals that one must model in order to effectively recognize and localize the group activities. Rather than focusing on directly recognizing group activities as most of the previous works do, we advocate the importance of introducing an intermediate representation for modeling groups of humans which we call *structured groups*. Such groups define the way people spatially interact with each other. People might be facing each other to talk, while others sit on a bench side by side, and some might stand alone. In this paper we contribute a method for identifying and localizing these structured groups in a single image despite their varying viewpoints, number of participants, and occlusions. We propose to learn an ensemble of discriminative interaction patterns to encode the relationships between people in 3D and introduce a novel efficient iterative augmentation algorithm for solving this complex inference problem. A nice byproduct of the inference scheme is an approximate 3D layout estimate of the structured groups in the scene. Finally, we contribute an extremely challenging new dataset that contains images each showing multiple people performing multiple activities. Extensive evaluation confirms our theoretical findings.

Keywords: Group discovery, Social interaction, Activity recognition

1 Introduction

In day-to-day environments we observe various types of complex group activities such as people conversing, waiting in a line, listening to a lecture, and eating together. Consider the images in Figure 1 - each shows multiple people involved in multiple different activities. To understand these scenes, we need to understand all of the activities, and to localize the activities we need to divide the people into groups with consistent spatial configurations consisting of individuals' poses, their relative poses, and their geometric patterns. For example, a group of people in conversation are often all sitting or standing while facing each other; a group of people standing in line may all face the same direction and stand one-behind-the-next; a group listening to a lecture might sit or stand while all face the same direction side-by-side. We call these consistent spatial configurations of people *structured groups*.

Structured groups are building blocks which can be composed to recognize and localize higher-level semantic group activities when combined with contextual information. For example, a line of people queuing plus two people at the front conversing plus a store context equals a group of people paying for something. Or a group of people sitting side-by-side plus a teacher at the front teaching plus a classroom context equals a lecture. Moreover, by identifying the members of a structured group, we can segment

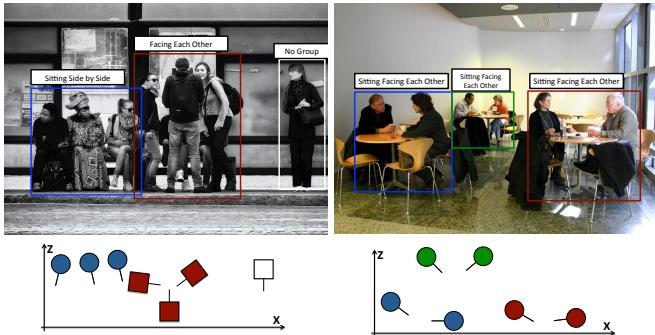


Fig. 1: Natural images often contain people forming multiple groups. Such groups can be categorized by the way people spatially interact with each other. People might be facing each other to talk, while others sit on a bench side by side, and some might stand alone. An image may contain multiple instances of the same group category. *Group discovery* is the problem of finding such *structured groups* despite their varying viewpoints, number of participants, and occlusions. In this paper we propose a framework for group discovery from a single image. Our approach is also capable of localizing the structure groups in 2D and half as shown in the lower part of the figure. Each person is depicted with a different color and symbol which corresponds to the structured group it belongs to (see figure 2 for a list of structure groups). Different colors corresponds to different group instances.

the people in a scene into different localized interactions. This will enable methods for identifying multiple activities in a given image.

In this paper, we focus on the problem of discovering structured groups of people and contribute a method for identifying and localizing these groups in a single image. We call this problem *group discovery*. Our approach seeks to 1) divide people into different classes of structured groups wherein participating individuals share the same patterns of interactions; 2) localize these groups in the 3D space; and 3) provide semantic descriptions to each structured group.

There are multiple challenges related to this problem domain. Detecting people in crowded scenes is extremely difficult due to occlusions and size variation. Also, once a person is found, it is still difficult to identify his/her pose and location. Moreover, given noisy detections but an unknown number of groups, the number of group assignments grows exponentially large. Finally, structured groups change their geometry and appearance because of viewpoint changes and other topological transformations.

To address these challenges, we contribute a method to segment people into semantically meaningful groups. We propose to use discriminative interaction patterns to encode the relationships between people in 3D. Capturing the relative poses and positions in 3D makes the approach viewpoint invariant. Importantly, the interaction patterns do not need to be supervised during training. As a key contribution, we propose an efficient *iterative augmentation* algorithm for solving the challenging inference problem. A nice byproduct of the inference scheme is an approximate 3D layout estimate of the structured groups in the scene. Finally, we contribute an extremely challenging new dataset that contains images each showing multiple people performing multiple activities.

1.1 Related Work

In the last decade, significant effort has been put toward understanding human activities at different levels of granularity. Several approaches have been proposed for classifying

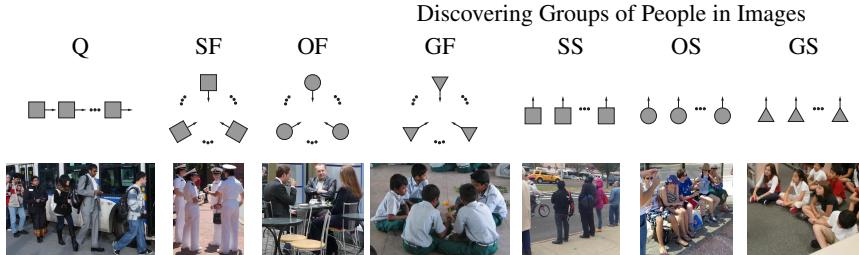


Fig. 2: The structured groups. Squares represent standing people, the circles are people sitting on chairs, and the triangles are people sitting on the ground. Arrows indicate individuals' orientations. See Sec. 2 for details.

the activity of a single person, including [21, 9, 24, 23]. Unfortunately, the activity of a single person in isolation is a poor indicator of the activity of a group of people.

Additional work has looked at the interactions between pairs of people [28, 36, 26, 35], and the collective activities of larger groups [6, 5, 19, 29, 16]. However most of these approaches perform image (or video) classification, assuming that there is only one activity. They do not localize the activities and hence have difficulty identifying multiple activities in the same scene. Recently, Desai and Ramanan [8] proposed a relational phraselet to model the relationship between two people (or generally two objects) and identify the poses, but the model is limited to a pair of people. Eichner and Ferrari [10] focus on human pose estimation in a group context, but they assume only a single group in the image. Pellegrini et al. [27] consider the problem of tracking people in crowded scenes. Using short tracks extracted from video data they jointly cluster people into groups and derive their longer trajectories. Leal-Taixe et al. [22] extend this idea further to incorporate more generalized interactions between people.

Recently, a number of methods were proposed to detect single person activities in video sequences [4, 13, 12]. Although they can localize multiple activities in the temporal direction, only one activity can be identified in one time slice. Odashima *et al.* [25] and Amer *et al.* [1] proposed methods to localize multiple collective activities, but they either leverage naive holistic models or evaluated only on datasets which contain limited view point and intraclass variability.

There is a long list of works addressing the problem of image segmentation [30, 32, 18]. These works take either individual pixels or superpixels as the basic unit, and segment the images into coherent regions. Unlike traditional bottom-up 2D pixel-based segmentation problems, the basic units in our work are human detections.

In this paper, we focus on discovering structured groups of people in images. These groups form the building blocks for detecting and classifying higher-level activities. As a contribution, we introduce the concept of *discriminative interaction* patterns that capture the characteristics of pairwise configurations. Unlike [26, 35, 5], which have a predefined set of pairwise interaction classes, we automatically discover an expressive dictionary of discriminative interactions. In Sec. 2, we describe our model representation that ties together structured groups, individuals and image observations. Sec. 3 explains our feature representation as well as the discriminative interactions. Secs. 4 and 5 describe how we solve the inference problem and learn the model parameters.

2 Model

The purpose of our model is to group together individual hypotheses into structured groups as they are observed in images (Fig. 3). We aim to localize and segment peo-

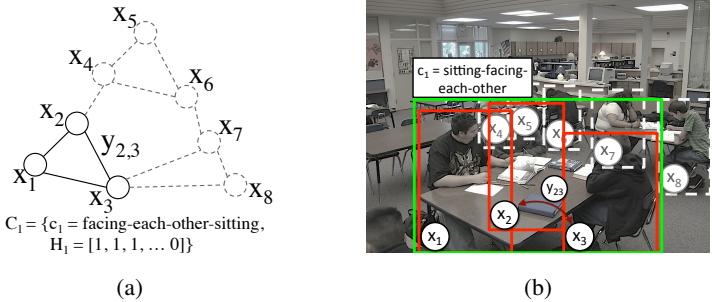


Fig. 3: An illustration of our model. Image I is represented by a set of individual detections \mathbb{X} . A group C_i (green bounding box) is represented by class c_i (sitting facing each other) and participating individuals (shown as solid black nodes in (a) and solid red bounding boxes in (b)) who are set to 1 in H_i . Individuals who are not group members are shown as dashed black nodes in (a) and white bounding boxes in (b). The group configuration potential is measured by the individual properties x_i and interaction $y_{i,j}$ between all participating individuals. (We model a fully connected graph, but omit edges in the figure for clarity.)

ple into 7 different types of structured groups. The categories are defined based on geometric relationships between participants. The first type is “*queuing*” (Q), a linear configuration of people, most of whom are facing the principal direction of the distribution. The second type is “*standing facing-each-other*” (SF), which is defined by a set of people in close proximity facing into a central point. Similarly, we define “*sitting facing-each-other*” (OF) and “*sitting on the ground facing-each-other*” (GF). Additionally, we define “*standing side-by-side*” (SS) that is characterized by a linear distribution of people with view point perpendicular to their distribution. Similarly, we introduce “*sitting side-by-side*” (OS) and “*sitting on the ground side-by-side*” (GS). These primitive types of structured groups provide useful information to understand social interaction among people. Examples of the structured groups are given in Fig. 2.

Let us define the set of individuals in an image I as $\mathbb{X} = \{x_1, x_2, \dots, x_n\}$. Each x_i encodes the properties of an individual. In this work we assume that a detector is available that can estimate each individual’s detection confidence, their pose, and their location in the image. The feature representation will be explained in detail in Section 3. Notice that the detector can be noisy and produce false alarm or missed detections, as well as poor localization and erroneous pose classification results.

Given a set of bounding boxes corresponding to person detections, we wish to identify the structured groups $\mathbb{C} = \{C_1, C_2, \dots, C_m\}$ in the image. Examples of structured groups are shown in Fig. 2. Each group description, C_k , consists of a binary membership vector $H_k = \{h_1^k, h_2^k, \dots, h_n^k\}$ indicating which individuals are in the group, and a class label c_k taking one of \mathcal{C} group labels. We also define a background group B that consists of the binary membership vector H_B indicating individuals that do not belong to any group. We assume that one individual can belong to at most one group, that is $\forall i, \sum_k h_i^k + h_i^B = 1$.

As an intermediate step between the individual detections and the group interactions, we model the interactions between pairs of people. During training, the detections, locations and poses of the people are given, however the interactions between pairs of people are not given and must be learned. Some examples of these learned in-

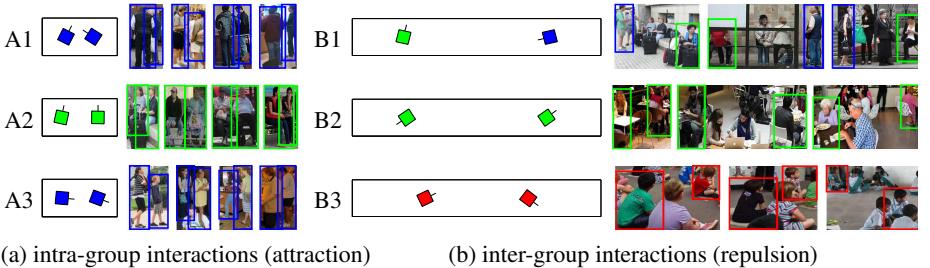


Fig. 4: A set of interaction examples learned from training data. The (a) and (b) show interactions learned from intra-group and inter-group pairs, respectively. Each row shows the 3D configuration of interactions from top view and the associated pairs in the training data. The pair examples show that our learned interactions are view-point invariant.

teractions are presented in Fig. 4. Let $\mathbb{Y} = \{y_{1,2}, \dots, y_{n-1,n}\}$ be the interaction variables that encode the relationships between pairs. Each y can have \mathcal{Y} discrete interaction labels that are learned from training data, as described in Sec. 3.4. These interaction labels encode different types of interactions that tend to occur within the same group instances (attraction interaction) or that tend to occur across different group instances (repulsion interactions) (See Fig. 4).

We now want to use this model to discover structured groups. Discovery is formulated in an energy minimization framework as follows:

$$\hat{\mathbb{C}} = \underset{\mathbb{C}}{\operatorname{argmin}} E(\mathbb{C}, \mathbb{X}, I) \quad (1)$$

where $\hat{\mathbb{C}}$ denotes the optimal set of structured groups and $E(\cdot)$ is an energy function. In order to capture both the characteristic pattern of each group as well as the compatibility between different groups, we define the energy function using a compositional model as follows (hereafter, we drop I for clarity):

$$E(\mathbb{C}, \mathbb{X}) = \min_{\mathbb{Y}} (\Psi_{XY}(\mathbb{X}, \mathbb{Y}) + \sum_i \Psi_{CX}(\mathbb{C}_i, \mathbb{X}, \mathbb{Y}) + \Psi_{BX}(\mathbb{B}, \mathbb{X}) + \Psi_{RY}(\mathbb{C}, \mathbb{Y})) \quad (2)$$

The first term Ψ_{XY} encodes the compatibility between interactions and individuals' properties, which is bottom-up information (Sec. 2.1). For example, in Fig. 1, two of the women sitting on the bench should be associated with an interaction like A2 in Fig. 4, but not A1. The second term Ψ_{CX} represents the *intragroup potential* which measures the compatibility of a set of individuals with a given structured group (Sec. 2.1). For instance, a “queuing” group (G_1) should be composed of individual detections with “standing” poses rather than “sitting” pose, and pairs of interactions similar to A3 in Fig. 4 (attractive interaction). The third potential is the background potential which measures the probability of a solitary person, while the last potential is the intergroup potential that measures the probability of two people begin in different groups (repulsive interaction). In the following subsections, we describe each potential function. We then discuss the details of the feature representation in Sec. 3.

2.1 Potentials

Bottom-up Interaction Potential The first potential encodes the compatibility between the properties of pairs of individuals (x_i, x_j) and the interaction variable $y_{i,j}$:

$$\Psi_{XY}(\mathbb{X}, \mathbb{Y}) = \sum_{i < j} w_{xy}^\top \psi_{xy}(x_i, x_j, y_{i,j}) \quad (3)$$

where the feature vector $\psi_{xy}(\cdot)$ encodes the observed interaction patterns between two people, and model parameters w_{xy} measure the compatibility between interaction patterns and a specific type of interaction (Sec. 3.4 and Fig. 5).

Intragroup Potential The second term encodes the intragroup potential. This reflects the characteristic patterns of each structured group as a function of individuals' appearances (Sec. 3.3) and their interaction variables. $\Psi(C_k, \mathbb{Y}, \mathbb{X})$ is designed to maximize the compatibility of individual's appearances and the relational patterns between individuals with a given type of structured group.

$$\Psi_{CXY}(C_k, \mathbb{Y}, \mathbb{X}) = \sum_i h_i^k w_{xc}^\top \psi_{xc}(x_i, c_k) + \sum_{i < j} h_i^k h_j^k w_{yc}^\top \psi_{yc}(y_{i,j}, c_k) \quad (4)$$

where w_{xc} and w_{yc} are model parameters, $\psi_{xc}(\cdot)$ encodes a person's appearance information (Sec. 3.3), and $\psi_{yc}(\cdot)$ is a feature vector encoding the co-occurrence of an interaction label and group label. The co-occurrence feature is a vectorized two dimensional matrix where only the specified elements have a value of 1, and others are 0.

Background Potential This potential captures individuals that do not belong to any group (background).

$$\Psi_{BX}(B, \mathbb{X}) = \sum_i h_i^B w_{xb}^\top \psi_{xb}(x_i) \quad (5)$$

where w_{xb} is the parameter and $\psi_{xb}(\cdot)$ encodes a person's appearance (Sec. 3.3).

Intergroup Potential This potential captures interactions of individuals belonging to different groups (repulsive interaction). $\Psi_{RY}(\mathbb{C}, \mathbb{Y})$ is formulated similarly to the intra-group potential.

$$\Psi_{RY}(\mathbb{C}, \mathbb{Y}) = \sum_{i < j} \left(1 - \sum_k h_i^k h_j^k \right) w_{yr}^\top \psi_{yr}(y_{i,j}) \quad (6)$$

where w_{yr} are the model parameters and $\psi_{yr}(\cdot)$ is a feature vector encoding the occurrence of a repulsive interaction label. Notice that unlike the intragroup potential, we accumulate all *ungrouped* pairs in Eq.6 that do not contribute to any intragroup potential in Eq.4. In turn, all the interaction labels either contribute to one group's intragroup potential (if attractive) or intergroup potential (if repulsive).

2.2 Model Characteristics

Our model has a number of favorable characteristics for discovering structured groups in images. First, it permits an arbitrary number of groups $\mathbb{C} = \{C_1, \dots, C_m\}$ in each image and provides a principled measurement of the groups' compatibility using both individuals' and interaction information. Secondly, it explicitly models whether an individual is a member of a group, so we can identify how people are clustered. Notice that we can even segment individuals participating in different instances of the same group type, which is very challenging when using distance-based clustering. This is enabled by the attractive and repulsive interaction model through the intragroup and intergroup potentials. Thirdly, our model also identifies which person detections are valid, which in turn improves the robustness of the group discovery despite noisy bottom-up detection results. Finally, the interaction variables $y_{i,j}$ enable us to share interaction patterns across different structured group categories, which would not be possible if the features were directly connected to the class \mathbb{C} without regard to \mathbb{Y} .

3 Feature Representation

In the previous section we described the potential functions that make up our model. Each of these potential functions relies on one or more feature representations. Here, we describe the details of those representations. For each image, we detect people and represent each detection with a pose confidence and the corresponding location of the person in 3D. Using this information, we encode the pairwise relationships using interaction features (Sec. 3.4) and the individuals’ contextual information using the unary group feature (Sec. 3.3).

3.1 Individual Pose Feature

Given an image, we first detect individual people using the Poselet detector [2]. We represent each detection hypothesis bounding box using a combination of the Poselet activation vector, MDP activation vector [33] and HOG descriptor [7]. Instead of using such a high dimensional vector directly to encode individual properties, we train SVM classifiers [3] equipped with histogram intersection kernel for individual pose classes and assign the confidence vector (probabilistic estimation) to individual feature p_i (e.g. we train person v.s. no person, standing v.s. sitting on an object v.s. sitting on the ground, and 8 viewpoints \times 3 poses classifiers). These individual pose vectors are used to represent the unary and interaction features.

3.2 3D Estimation

In order to robustly encode the relationship between people in images, we propose to estimate each person in 3D using a technique similar to [14]. Each person z_i is parametrized by their 3D location, height, and pose. The camera Θ is represented by its focal length, pitch angle, and height (we assume zero yaw and roll angles). With the assumption that every person is located on the same ground plane, the camera parameter Θ and the presented human $Z = \{z_1, \dots, z_n\}$ are estimated by optimizing the function:

$$E(\Theta, Z, I) = \omega_{\Theta I} \Psi(\Theta, I) + \omega_{\Theta Z} \Psi(\Theta, Z) + \omega_{Z I} \Psi(Z, I) + \omega_{\Theta} \Psi(\Theta). \quad (7)$$

The first term captures the compatibility between the camera parameter Θ and the image feature. The second term captures how well the humans in configuration H fit into the scene given the camera parameter Θ . The third term captures how likely the human configuration H is given the observed image. The last term accounts for the prior on Θ . Given the function E , we solve Θ and Z in a two-step fashion. We first solve Θ using the RANSAC algorithm, which iterates between 1) sampling three people and their pose from the detection set and 2) computing the sample score E by optimizing Θ given the sampled people. Once we generate enough samples, the optimized camera parameter Θ is obtained from the sample with the highest score E . Finally, we compute Z (locations, heights, and poses of all human) by maximizing E given Θ .

3.3 Individual Unary Feature

Inspired by [6, 20], we represent an individual’s “appearance” information with a contextual descriptor that captures the relative distribution of other people in the scene. Similarly to [6], we discretize the space around the individual into multiple radial and

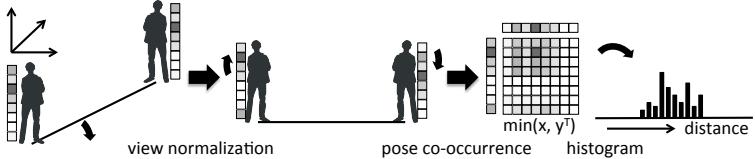


Fig. 5: We obtain the interaction feature via 1) view normalization, 2) pose co-occurrence encoding, 3) semantic compression and 4) spatial histogram. Given the 3D coordinates of two detections, the location of individuals are transformed to a canonical view and the corresponding pose features are shifted accordingly. Subsequently, the co-occurrence pattern is encoded by taking minimum of the transformed pose confidence values. The elements that deliver the same semantic concepts are accumulated to reduce the feature dimension and provide semantic invariance. Finally, the compressed representation is assigned to the spatial bin corresponding to the distance between the two individuals.

angular bins and pool the pose confidence vectors of other detections. In order to make the descriptor robust under noisy detections, two nearest bin centers are obtained for each dimension (angular and radial) and the pose confidence vectors are assigned with linear interpolation weights. We use 8 angular bin centers and 3 radial bin centers.

Given the contextual descriptor for each individual, we represent the unary feature $\phi_x(x_i)$ using the confidence value of SVM classifier trained on the group categories. In order to deal with outliers (NA) and false positive (FP) detections, we add two more categories to the category set. With $\phi_x(x_i)$, $\psi_{xc}(x_i, c_k)$ (Eq. 4) are obtained by shifting $|\mathcal{C}| * c_k$ dimensions to make it compatible with the parameters in w_{xc} , where \mathcal{C} is the set of structured group categories including NA and FP.

3.4 Pair Interaction Feature

The interaction pattern between two individuals is encoded by a spatial descriptor with view invariant relative pose encoding. Given the 3D locations of two individual detections z_i, z_j and two pose features p_i, p_j , we represent the pairwise relationship using view normalization, pose co-occurrence encoding, semantic compression and a spatial histogram (see Fig. 5 for illustration).

The view normalization is performed by rotating the two people in 3D space by θ with respect to their midpoint, making their connecting line perpendicular to the camera view point. In this step, the pose features are also shifted accordingly (e.g. if $\theta = 45^\circ$, shift 1 dimension with a cycle). Then, the co-occurrence feature is obtained by building a 2-dimensional matrix in which each element (r, c) corresponds to $\min(p_i(r), p_j(c))$. Although the feature is view invariant, there are still elements in the matrix that deliver the same semantic concepts (e.g. left-left and right-right). To reduce such unnecessary variance and obtain a compact representation, we perform another transformation by multiplying a semantic compression matrix Sc to the vector form of the co-occurrence feature. The matrix Sc is learned offline by enumerating all possible configurations of view points and grouping the pairs that are equivalent when rotated by 180 degrees. Finally, we obtain the pair interaction descriptor by building a spatial histogram based on the 3D distance between the two (bin centers at 0.2, 0.6, 2.0 and 6.5 m). Here, we use linear interpolation similarly to contextual feature in Sec. 3.3. Given the interac-

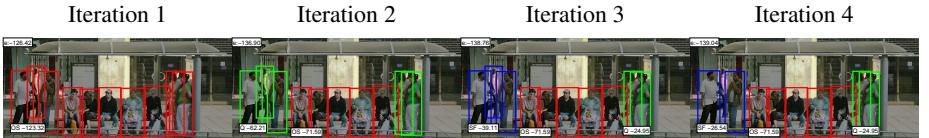


Fig. 6: Illustration of the iterative augmentation algorithm. We show the different groups discovered in each iteration with different colored bounding boxes. The number on top-left corner shows overall energy value given the configuration and numbers on bottom-left corner of each group shows the energy contribution of each group. Notice that the algorithm tends to group larger people in one iteration since a larger incorrect group minimizes overall objective value better than a smaller but correct group. Our algorithm is capable of fixing such mistakes made in earlier iterations and arrives at an optimal solution in a few iterations.

tion descriptor for each pair, we represent the interaction feature $\phi_{xx}(x_i, x_j)$ using the confidence value from an SVM classifier trained on a dictionary of interaction labels \mathcal{Y} .

Each interaction label is defined as a characteristic 3D spatial relationship between a pair of people. The key property is that each interaction label must be view-invariant, so that we can propagate a view invariant signal to the group model using $\psi_{yc}(y_{i,j}, c_k)$. To achieve view invariance, we obtain a set of interaction labels by agglomeratively clustering all possible pairs of individuals using a similarity metric S defined below. Given 3D location as well as poses of a pair z_i, z_j , we first align each of them by rotating along the center, so the $\theta = 0$. The similarity metric S between (z_i, z_j) and (z_k, z_l) is defined as follows:

$$S((z_i, z_j), (z_k, z_l)) = \lambda(d_{ij} - d_{kl})^2 + \gamma[\Delta(\theta_i, \theta_k)^2 + \Delta(\theta_j, \theta_l)^2] \quad (8)$$

where d_{ij} is the distance between z_i and z_j and θ_i is the individual pose angles of z_i after the alignment. $\Delta(\theta_i, \theta_k)$ computes the difference between θ_i and θ_k . The weights λ and γ are empirically set to 2.5 and 1. Any clusters containing more than T instances are kept as possible interaction labels. In practice, we set $T = 10$. In order to learn both intragroup and intergroup interaction labels, we perform the dictionary learning separately for the pairs observed in the same groups and the pairs observed across different groups. Some examples of learned interaction labels are shown in Fig. 4.

Given the SVM interaction classifier confidences vector ϕ_{xx} , the $\psi_{xy}(x_i, x_j, y_{i,j})$ (Eq.3) are obtained by taking the confidence value of the interaction label $y_{i,j}$.

4 Inference

Given the potentials and features defined thus far, we can model the interactions between people in an image. To find the optimal set of groups $\hat{\mathbb{C}}$ that minimize equation 1, we need to perform inference on this model. The inference problem, however, is very challenging since there are an unknown number of groups \mathbb{C} with different group types in a scene. To cope with the challenges, we propose the *iterative-augmentation* algorithm.

Let us define the group augmentation operator \oplus as $\mathbb{C}_k \oplus C_{new} = \mathbb{C}_k^- \cup C_{new}$ where $\forall i, k, h_i^{k-} = 0$ if $h_i^{new} = 1$, $h_i^{k-} = h_i^k$ otherwise (h_i^k represents the membership of individuals to a group, Sec. 2). Clearly, the operator is not commutative and gives priority for the new group to *include* individuals as participants while keeping the assignment constraint $\sum_h h_i^k + h_i^B = 1$.

The iterative augmentation algorithm starts the inference with an empty set of groups $\hat{\mathbb{C}}_0 = \emptyset$. Then, in each iteration we find a single group that minimize the function:

$$\nabla \Psi(C_k; \hat{\mathbb{C}}_{k-1}, \mathbb{X}, \mathbb{Y}) = \Psi(\hat{\mathbb{C}}_{k-1} \oplus C_k, \mathbb{X}, \mathbb{Y}) - \Psi(\hat{\mathbb{C}}_{k-1}, \mathbb{X}, \mathbb{Y}) \quad (9)$$

The new group $\hat{\mathbb{C}}_k$ is obtained by $\hat{\mathbb{C}}_{k-1} \oplus C_k$. The key property of this algorithm is that it can fix mistakes made in early iterations (see Fig. 6) and can find the optimal solution if it finds optimal C_k for Eq. 9. We optimize Eq. 9 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:

$$\nabla \Psi(\hat{H}_k) = \sum_i \psi_u(\hat{h}_i^k) + \sum_{i < j} \psi_p(\hat{h}_i^k, \hat{h}_j^k) \quad (10)$$

Since this CRF is fully connected and contains high order cliques, it is not possible to obtain globally optimal solution in an efficient way. Instead, the energy function is approximated with a mean field distribution $Q(\hat{H}_k) = \prod_i Q(\hat{h}_i^k)$ that minimizes the KL divergence between Q and $P(\hat{H}_k) = \frac{1}{Z} \exp(-\nabla \Psi(\hat{H}_k))$ using mean-field message passing (MFMP) [17]. Given the approximate distribution Q , the optimal solution is found by taking the maximizing state of each variable in the marginal distribution. Please see the supplemental material for details of the reformulation and the derivation of the mean-field message passing algorithm. In practice, we run both the MFMP algorithm and the greedy algorithm and take the solution gives better objective value.¹ Although the interaction labels \mathbb{Y} could be optimized jointly in this algorithm by incorporating a high order potential, in practice we obtain \mathbb{Y} separately using the bottom-up signal for computational reasons. Computing the joint model efficiently is future work.

5 Model Learning

To train our model, we need to learn the parameters w that maximize its discriminative power. We learn these from training data that contains supervised annotations of 1) bounding boxes around individual people, 2) pose labels for individual people, and 3) the group annotations. Group annotations are done by labeling each person to his associated group instances, or to not belonging to any groups. For instance, in Fig. 1 (right), the man in the gray jacket and the women in front of him belong to a “sitting-facing-each-other” group while the others are belong to different instances of the same group type.

Given the interaction dictionary \mathcal{Y} (Sec. 3.4) and a set of training images $\{\mathbb{X}_i, \mathbb{C}_i\}_i$, the model parameters are learned using the structured support vector machine framework [15]. In order to obtain the group association \mathbb{C}_i and interaction labels \mathbb{Y}_i , we find the optimal association between detections \mathbb{X}_i and ground truth human annotations by computing the intersection-over-union (IOU) between the two. If the IOU is larger than 0.5, we transfer the ground truth group association and interaction labels to the corresponding detections and pairs of detections. If not, we assign a false positive group and

¹ We observe that two algorithms are complementary to each other in the experiments. Among all the experiments, MFMP achieves a better solution (with smaller objective values) for 41% of the times, while greedy win for 33% of the times.

	Pose			Interaction	Group Context
# classes	2	4	25	59	9
Train	83.1 / 83.6	78.0 / 77.7	42.5 / 47.6	34.1 / 48.1	52.6 / 66.7
Test	79.6 / 80.5	71.0 / 66.9	41.7 / 33.6	28.8 / 35.2	43.0 / 52.7

Table 1: Classification accuracy of input features. Poselet detections are used in this experiment. We show both average per class accuracy and overall accuracy as the data is unbalanced.

a false positive interaction label to any pair that is connected to the detection. Then, using the complete set of information $\{\mathbb{X}_i, \mathbb{Y}_i, \mathbb{C}_i\}_i$, we obtain the corresponding model parameters w using the following structural SVM formulation:

$$\min_{w, \xi} \frac{1}{2} \|w\|^2 + C \sum_i \xi^i, \text{ s.t. } \Psi_{\mathcal{Y}, w}(\mathbb{C}, \mathbb{X}_i, \mathbb{Y}_i) - \Psi_{\mathcal{Y}, w}(\mathbb{C}_i, \mathbb{X}_i, \mathbb{Y}_i) \leq \xi^i - \delta(\mathbb{C}, \mathbb{C}_i), \forall i, \mathbb{C}$$
(11)

where C is a hyper parameter in an SVM and ξ^i are slack variables. $\delta(\mathbb{C}, \mathbb{C}_i)$ represents the loss function of a discovered group \mathbb{C} . More details of the learning procedure can be found in the supplemental materials.

6 Experimental Evaluation

Dataset: We test our algorithm on a newly proposed challenging dataset, *Structured Group Dataset* (SGD)². The dataset contains 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. Examples of the structured groups are shown in the Fig. 2. We contribute supervised annotations of individual person properties - bounding boxes, individual poses (standing, sitting on an object, sitting on the floor), and 8 different view points (front, front-left, ..., back-right). We also provide annotations of all groups. Please see the supplemental material for examples and statistics.

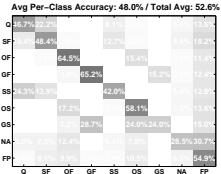
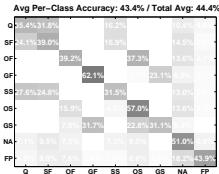
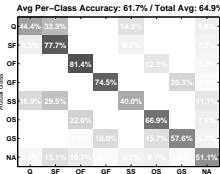
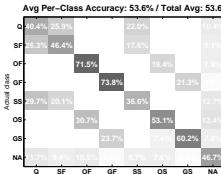
Experimental Setup: In order to provide an extensive evaluation of the method, we run 5 fold training and testing over the entire dataset. In each split, we learn the pose classifier, the interaction dictionary and classifier, and contextual group classifier in sequential order. To prevent overfitting we train each low level classifier using 10 fold cross validation over the training set and assign the classifier output in each fold separately. The cross-validated confidence values are used as features for model learning. The entire training data are used to provide features for the testing data. Please see the Tab. 1 for the classification accuracies for the first set of training and testing splits.

Evaluation Metric: Evaluating structured groups is complex. To evaluate whether a group is correct we first need to determine if the individuals were detected correctly and if they were, whether they belong in the same group. Assigning individual detections to ground truth detections must be done carefully to provide a fair evaluation. For example, there could be many detection boxes overlapping with a ground truth annotation. Evaluating the grouping is also complex as accidentally dividing or merging clusters can vastly affect the outcome [34].

In the following experiments, we report two metrics. First, we report the precision, recall, and corresponding F1-measure ($2PR/(P + R)$) value of our solution. Group

² The dataset and our code are available at <http://cvgl.stanford.edu/projects/groupdiscovery/>.

GT detections



Context Classifier

Our full model

Context Classifier

Our full model

Fig. 7: Confusion tables for group category classification with GT and poselet detections. We compare the accuracy of the output of context feature alone (Sec. 3.3) and the final output of our model. In both GT and poselet detection experiments, we observe significant improvements.

detections are evaluated using the intersection over union ratio of the group participants. This is computed by dividing the number of individuals that are both included in a ground truth group and a discovered group (intersection) by the number of individuals that are either participating in the ground truth group or the discovered group (union). If the ratio is larger than 0.5, we declare the discovered group a true positive. Following the PASCAL evaluation convention [11], only one group is associated with a ground truth group. This measure can evaluate the segmentation accuracy of individuals into groups as well as the categorization accuracy of each discovered group. When evaluating the group discovery with poselet detections, we ignore the ground truth annotations that do not match with any detection hypothesis to separate the error caused by detector itself. Second, we show the individual person labeling accuracy in a confusion table to evaluate the assignment of individuals to semantic groups.

Baseline Method: As a baseline for our evaluation, we propose to group objects based on proximity. A similarity matrix is constructed based on the estimated 3D distance between people and Normalized Cuts [31] is used to cluster the people into k groups. NCuts requires k to be given (denoting NCut+K), so we provide the correct ground truth number of groups in each image (including groups for outliers and false positive detections). This is a very important piece of information that our system does not have, so this baseline is very strong. The group category label is assigned by taking the votes of individual contextual classifier outputs (Sec. 3.3) and no interaction labels are considered here.

Component Evaluations: To better understand the contributions of the components of our system, we also compare against multiple reduced versions of our model as follows.

- (w/out interactions) We exclude the intragroup and intergroup interactions by replacing the pairwise edges in our graph with conventional distance-based edges. Interaction labels are replaced by 4 distance labels corresponding to the distance bins used in the pair interaction feature (Sec. 3.4).
- (w/out 3D) We use only 2D information instead of the estimated 3D information when computing the interaction feature. The distance between people is normalized by their average width.

We also provide all results on both detected persons as given by the Poselet detector [2] (Poselet) and on the ground truth person detections (GT) to separate the system performance from the performance of the person detection algorithm used.

Results and Analysis: We start with the precision, recall, and F1 measure comparison. Table 2 (top) summarizes the results for all of the baseline and reduced methods given

ground truth human annotation person detections. Table 2 (bottom) gives the same results given the *Poselet human detections*.

We observe that there is a large gap between the group discovery results obtained using GT and poselet detections. Actual detection outputs are often noisy in terms of localization, missing detections and false positives. This makes high-level reasoning extremely challenging. Nevertheless, our algorithm shows robust results as seen in qualitative examples shown in Fig. 8 and the quantitative metrics shown in the Tab. 2.

We also notice that conventional clustering (NCuts) does not perform well even though the ground truth number of clusters are given and the same unary features are provided for the categorization. This confirms that the data set is extremely difficult and a more complex approach is required.

Let us consider the partial models. The partial model without pairwise interactions does not perform as well as the full model. This implies that reasoning about the interaction between pairs of people is critical for finding groups.

The 2D-based model achieves much lower accuracy than the 3D-based model. This is not surprising given that the 2D model is not view or distance invariant.

Our experiments also show that the algorithm can effectively estimate the number of groups present in the scene. We found that the mean absolute error of the predicted number of groups per image is 0.59 when GT detections are used, and 0.71 when poselet detections are used. The mean true number of groups is 4.25 for entire test set.

Finally we study the accuracy of classifying the group category for each individual, with results in the confusion tables in Fig. 7. For each individual, the category label is derived from the label of its group. The result suggest that reasoning with the full model

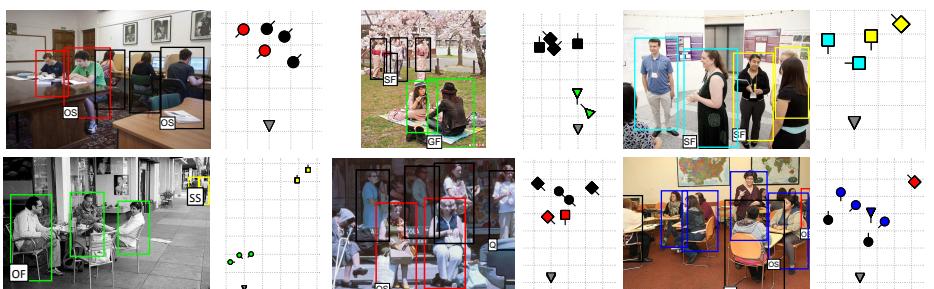


Fig. 8: Qualitative examples of the results obtained using our full model with poselet detections. We show the image configuration of groups on the left and corresponding 3D configuration on the right. Different colors represent different groups, the type of each structured group is overlaid on the bottom-left 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). Detections that are not assigned to outlier (NA), false positive (FP) or background are not shown to avoid clutter.

Method	P/R	Q	SF	OF	GF	SS	OS	GS	Avg
Group Discovery with Ground Truth Detections									
NCuts+K	Prec	12.29	47.80	50.13	43.55	25.94	47.25	43.03	38.57
	Recall	35.50	40.91	76.73	76.00	25.23	39.21	30.87	46.35
	F1	18.26	44.09	60.64	55.37	25.58	42.86	35.95	42.10
w/o interactions	Prec	6.76	39.48	44.88	56.34	21.68	40.08	21.89	33.01
	Recall	1.91	55.06	56.52	60.00	14.22	38.96	19.13	35.11
	F1	2.98	45.99	50.03	58.11	17.17	39.51	20.42	34.03
w/o 3D	Prec	32.29	44.96	46.98	52.23	36.51	38.18	26.63	39.68
	Recall	11.83	57.92	52.66	58.50	15.83	36.29	19.57	36.09
	F1	17.32	50.62	49.66	55.19	22.08	37.21	22.56	37.80
full model	Prec	41.86	55.78	62.48	60.19	39.08	53.85	37.65	50.13
	Recall	27.48	64.55	65.56	65.00	21.33	40.86	26.52	44.47
	F1	33.18	59.85	63.98	62.50	27.60	46.46	31.12	47.13
Group Discovery with Poselet Detections									
NCuts+K	Prec	6.47	11.81	16.58	19.38	13.52	18.00	8.39	13.45
	Recall	29.03	34.96	34.19	53.39	24.37	44.95	12.00	33.27
	F1	10.58	17.66	22.33	28.44	17.39	25.71	9.88	19.16
w/o interactions	Prec	9.32	18.25	20.49	29.80	16.37	20.72	11.11	18.01
	Recall	8.87	30.49	36.25	50.00	14.21	39.63	10.00	27.06
	F1	9.09	22.83	26.18	37.34	15.21	27.21	10.53	21.63
w/o 3D	Prec	16.95	21.13	26.97	34.90	25.14	20.77	9.84	22.24
	Recall	16.13	28.86	43.19	56.78	22.34	37.50	6.00	30.11
	F1	16.53	24.40	33.21	43.23	23.66	26.73	7.45	25.58
full model	Prec	25.74	26.40	30.61	36.21	30.57	23.01	13.19	26.53
	Recall	28.23	42.28	50.13	53.39	24.37	33.78	12.00	34.88
	F1	26.93	32.50	38.01	43.15	27.12	27.37	12.57	30.14

Table 2: Precision, recall, and F1 measure given ground truth detections (top) and Poselet detections (bottom). Each column shows the precision and recall for each structured group category. Our model shows clear improvement over the baseline NCut+K and partial models. In some cases NCut+K does better than our model, but note that we provided the ground truth number of groups to NCut algorithm that is not available in practice. The best numbers in each column are bold.

improves individual classification over the baseline unary classifier that only looks at local context. We show results for both ground truth and detected persons.

7 Conclusion

In this paper, we tackled a new challenging problem, *group discovery* in images. As a key contribution, we introduced the concept of discriminative interaction patterns and proposed a view invariant interaction feature to robustly encode the patterns. The interaction patterns enabled us to segment different instances of groups properly. Our *iterative augmentation* algorithm effectively found the number of structured groups in images and identified their participants. We demonstrated the effectiveness of our algorithm using both quantitative and qualitative experimental results on a new and difficult dataset. This approach is now ready to be used as a step toward higher-level activity understanding.

Acknowledgement: The work is partially supported by an ONR award N000141110389.

References

1. Amer, M.R., Xie, D., Zhao, M., Todorovic, S., Zhu, S.C.: Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In: ECCV (2012)
2. 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>
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
4. Chen, C.Y., Grauman, K.: Efficient activity detection with max-subgraph search. In: CVPR (2012)
5. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: ECCV (2012)
6. Choi, W., Shahid, K., Savarese, S.: What are they doing? : Collective activity classification using spatio-temporal relationship among people. In: VWS (2009)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
8. Desai, C., Ramanan, D.: Detecting actions, poses, and objects with relational phraselets. In: ECCV (2012)
9. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (2005)
10. Eichner, M., Ferrari, V.: We are family: Joint pose estimation of multiple persons. In: ECCV (2010)
11. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>
12. Hoai, M., De la Torre, F.: Max-margin early event detectors. In: CVPR (2012)
13. Hoai, M., Lan, Z.Z., De la Torre, F.: Joint segmentation and classification of human actions in video. In: CVPR (2011)
14. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. IJCV (2008)
15. Joachims, T., Finley, T., Yu, C.N.: Cutting-plane training of structural svms. Machine Learning (2009)
16. Khamis, S., Morariu, V.I., Davis, L.S.: Combining per-frame and per-track cues for multi-person action recognition. In: European Conference on Computer Vision (2012)
17. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
18. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.S.: Graph cut based inference with co-occurrence statistics. In: ECCV (2010)
19. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: NIPS (2010)
20. Lan, T., Wang, Y., Mori, G., Robinovitch, S.: Retrieving actions in group contexts. In: International Workshop on Sign Gesture Activity (2010)
21. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV (2003)
22. Leal-Taixe, L., Fenzi, M., Kuznetsova, A., Rosenhahn, B., Savarese, S.: Learning an image-based motion context for multiple people tracking. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
23. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos “in the wild”. In: CVPR (2009)
24. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. IJCV (2008)

25. Odashima, S., Shimosaka, M., Kaneko, T., Fukui, R., Sato, T.: Collective activity localization with contextual spatial pyramid. In: ECCV Workshops and Demonstrations (2012)
26. Patron-Perez, A., Marszałek, M., Zisserman, A., Reid, I.D.: High five: Recognising human interactions in TV shows. In: BMVC (2010)
27. Pellegrini, S., Ess, A., Gool, L.V.: Improving data association by joint modeling of pedestrian trajectories and groupings. In: European Conference on Computer Vision (ECCV) (2010)
28. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: ICCV (2009)
29. Ryoo, M.S., Aggarwal, J.K.: Stochastic representation and recognition of high-level group activities. IJCV (2010)
30. Shi, J., Malik, J.: Normalized cuts and image segmentation. PAMI (2000)
31. Shi, J., Malik, J.: Normalized cuts and image segmentation. PAMI 22(8), 888–905 (2000)
32. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. IJCV (2009)
33. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: European Conference on Computer Vision (2012), <http://arxiv.org/abs/1205.3137>
34. Ullnikrishnan, R., Pantofaru, C., Hebert, M.: Toward objective evaluation of image segmentation algorithms. PAMI 29(6), 929–944 (2007)
35. Yang, Y., Baker, S., Kannan, A., Ramanan, D.: Recognizing proxemics in personal photos. In: CVPR (2012)
36. Yao, A., Gall, J., Van Gool, L.: A hough transform-based voting framework for action recognition. In: CVPR (june 2010)