3D Visual Proxemics: Recognizing Human Interactions in 3D from a Single Image

Ishani Chakraborty Hui Cheng Omar Javed
SRI International, Princeton, NJ 08540
ishani.chakraborty,hui.cheng,omar.javed@sri.com

Abstract

We present a unified framework for detecting and classifying people interactions in unconstrained user generated images. [1] *Unlike previous approaches that directly map people/face locations in 2D image space into features for classification, we first estimate camera viewpoint and people positions in 3D space and then extract spatial configuration features from explicit 3D people positions. This approach has several advantages. First, it can accurately estimate relative distances and orientations between people in 3D. Second, it encodes spatial arrangements of people into a richer set of shape descriptors than afforded in 2D. Our 3D shape descriptors are invariant to camera pose variations often seen in web images and videos. The proposed approach also estimates camera pose and uses it to capture the intent of the photo. To achieve accurate 3D people layout estimation, we develop an algorithm that robustly fuses semantic constraints about human interpositions into a linear camera model. This enables our model to handle large variations in people size, heights (e.g. age) and poses. An accurate 3D layout also allows us to construct features informed by Proxemics that improves our semantic classification. To characterize the human interaction space, we introduce visual proxemes; a set of prototypical patterns that represent commonly occurring social interactions in events. We train a discriminative classifier that classifies 3D arrangements of people into visual proxemes and quantitatively evaluate the performance on a large, challenging dataset.*

a. Group interaction b. Family photo c. Group photo d. Couple and audience e. Crowd f. Speaker and audience

Figure 1: People configurations and the camera-person's perspective provide strong cues about the type of social interaction that the people are participating in. The proposed method uses 2D face locations from a single image to estimate the camera pose and the spatial arrangement of people in 3D.

1. Introduction

A significant number of images and videos uploaded to the Internet, such as YouTube videos or Flickr images, contain scenes of people interacting with people. Studying people interactions by analyzing their spatial configuration, also known as Proxemics in anthropology, is an important step towards understanding web images and videos. However, recognizing human spatial configurations (i.e., proxemes) has received relatively little attention in computer vision, especially for unconstrained user generated content.

Figure 1 shows six typical types of people interactions that are often seen in Internet images and video frames: They are (1) Group Interaction, (2) Family photo, (3) Group photo, (4) Couple with an audience, (5) Crowd, and (6) Speaker with audience. From these images, it is important to note that the people configurations in the 3D space would better reflect the type of interaction than the configurations in a 2D image space. For example, Figure 1(a), (d), (e) and (f) all have many faces distributed throughout the image space, but they have very different spatial arrangements that can be distinguished in the 3D

[1]This work has been supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11-PC20066. The U.S. Government is au- thorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and con- clusions contained herein are those of the authors and should not be in- terpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Govern- ment.

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.437

space. Additionally, not only how people are organized spatially, but also how the shots are framed (i.e. the relative camera location, direction and pose) convey the type of proxemes depicted in these images. For example, in order to capture the whole group and to avoid occlusion, high-angle shots are used for group interaction (Figure 1(a)) and crowd (e). On the other hand, to capture the focus of attention, or principals in an event, such as a family portrait (Figure 1(b)), couples in a ceremony (Figure 1(d)) and speaker with an audience (Figure 1(f)), eye level shots are used. For artistic impression and better capture of the people in foreground without concerns of occluding the background, low-angle shots are used, especially for group photos as shown in Figure 1(c).

A number of research groups [19, 5, 10, 9] have conducted insightful studies for understanding people interactions in images and videos, though with limited scope. Most of these approaches [19, 5] perform their analysis in the 2D camera space. Although these approaches demonstrated their effectiveness, their robustness is fundamentally limited by the 2D analysis paradigm and cannot handle the diversity in camera pose and people depths often seen in user generated Internet content.

In recent works, [10] proposes to estimate 3D location of people using faces and use these locations to detect social interaction among people. In [9], locations of faces in the 3D space around a camera wearing person are used to detect attention patterns. However, these approaches only attempt to detect a very limited set of human interactions and their 3D estimation cannot effectively handle the diversity of people in terms of age (big adults vs. small children), height (tall vs. short), and the diversity of peoples poses such as sitting, standing and standing on platforms. Additionally, these approaches do not take camera location and pose into account when analyzing people interactions, which can be an important clue about the intent of the shot.

The theory of Proxemics [11] studies the correlation between human's use of space (proxemic behavior) and interpersonal communication. It provides a platform to understand the cues that are relevant in human interactions. Proxemics has been applied in the field of cinematography where it is used for optimizing the scene layout and the position of the camera with respect to the characters in the scene. We believe these concepts are relevant beyond cinematic visuals and pervade all types of images and videos captured by people. Inspired by the role of Proxemics in visual domain, we propose to analyze and recognize human interactions using the attributes studied in this theory.

In this paper, we propose a unified framework called 3D Visual Proxemics Analysis (VPA3D), for detecting and classifying people interactions from a single image. VPA3D first estimates people/face depths in 3D, then performs perspective rectification to map people locations from the scene space to the 3D space. Finally, a set of spatial and structural features are used to detect and recognize the six types of people interaction classes.

The proposed VPA3D approach surpasses state-of-the-art people configuration analysis in the following three aspects. First, VPA3D uses *3D reasoning for robust depth estimation* in the presence of age, size, height and human pose variation in a single image. Second, a set of *shape descriptors derived from the attributes of Proxemics* is used to capture type of people interaction in the eyes of each individual participant not only for robust classification but also for classification of individuals role in a visual proxeme. Additionally, the *types of camera pose* are used as a prior indicating possible intent of the camera-person who took the picture. Third, to characterize the human interaction space, we introduce *visual proxemes; a set of prototypical patterns* that represent commonly occurring people interactions in social events. The source of our visual proxemes is the NIST TRECVID Multimedia Event Detection dataset [2] which contains annotated data for 15 high-level events. A set of 6 commonly occurring visual proxemes (shown in Figure 1) are selected from keyframes containing groups of people. We train a discriminative classifier that classifies 3D arrangements of people into these visual proxemes and quantitatively evaluate the performance on this large, challenging dataset.

2. Related Work

Group dynamics is studied for visual surveillance, anomaly detection and in smart environments. Social force model was proposed in [18] to understand pedestrian behavior and interaction energies between people tracks was used in [8] to detect abnormal and violent interactions. Crowd context is exploited to understand collective activities e.g., "queuing" and "talking" in [6]. Learning patterns of crowd or group behavior in time and space requires long-term and accurate tracking and identity associations of people's motions in videos.

Proxemics is a subfield of anthropology that involves study of people configurations [11]. Its core idea that spatial configuration among people is strongly related to social interactions has been recently adopted by the computer vision community. Most of these works consider face detections to localize people in images. There are two main directions in this line of work - *2D approaches* that directly translate image based features into concepts, and *3D approaches* that translate detected faces into the 3D scene and then derive features from the 3D layout. Examples of 2D approaches include [5] in which authors predict pairwise relationships e.g., couple, sibling etc. from facial attributes and face subgraphs, and [19], in which authors label pairs of people with a physically-based touch codes such as Hand-hand, Hand-torso etc. These works primarily focus on mining pairwise relationships in personal photos.

3D layout has been mostly considered for retrieving the gaze directions and attention patterns in groups of people. These methods use rough estimates of 3D locations determined from face size variations to seed the directions of gazelines. In [17], by detecting 3D gaze volumes, the authors find if people are looking at each other. Gaze direction is combined with first-person movement in an MRF model in [9] to predict social interactions like "dialogue" and "discussion". Recently, Gallagher et al. [10] have presented a systematic study for understanding images of groups of people where they look into 3D spatial structure by recovering camera parameters from face locations. However, their method for camera calibration does not allow variations in people poses and camera viewpoints.

Camera calibration from single view image was traditionally performed by analyzing vanishing points [7, 16]. These algorithms estimate vanishing points based on camera motion constraints in [16] and rigidity constraints of architectural scenes in [7]. Recently, Hoiem et al. [12] have proposed an approximate camera calibration model that assumes grounded objects of known heights and restricted intrinsic camera parameters. They provide an algebraic solution that jointly estimates the horizon line and the object depths. Our perspective rectification

model adds robustness to this approach and combines it with semantically derived constraints. Semantic constraints have previously been explored for relative depth ordering of textured regions in [15]. In contrast, we introduce constraints to estimate perspective-corrected, explicit positions of faces in 3D space as well as the camera height relative to the faces.

3. 3D Visual Proxemic Analysis: Overview

Broadly, our 3D Visual Proxemic Analysis formulates a framework that unifies three related aspects, as illustrated in the system pipeline (Figure 2). First, we introduce **Visual Proxemics** as a prior domain knowledge that guides our analysis and recognition of human interactions in images and videos. Then, we describe a novel **Perspective Rectification** algorithm to estimate people/face depths in 3D and camera view from face detections in images. Finally, we categorize images into common types of social interactions (i.e., proxemes) in the **Visual Proxeme Classification** stage by combining the estimates of face positions and camera view with our knowledge of Visual Proxemics through spatial and structural features in the 3D space.

3.1. Visual Proxemics Description

Proxemics is a branch of cultural anthropology that studies man's use of space as a way for nonverbal communication [11]. In this work, we leverage the findings in Proxemics to guide us in our analysis and recognition of human interactions in visual media including images and videos. We call this Visual Proxemics and summarize our taxonomy of attributes in Figure 3.

A key concept in Proxemics is "personal space" that associates **inter-person distance** with the relationships among people. It is categorized into four classes: "intimate distance" for close family, "personal distance" for friends, "social" distance for acquaintances and "public distance" for strangers. Additionally, people configuration needs to support the communicative factors such as physical contacts, touching, visual, and voice factors needed in an interaction [1]. Based on these factors, we can see that certain types of the interactions will result in distinct **shape configurations** in 3D. For example, in Figure 1a, to enable direct eye contact between any pair of participants in a group interaction, people align themselves in a semi-circular shape. In contrast, if two people are the focus of attention, as in Figure 1d, we have **multiple shape layers**, where the two people at the center of attention share an intimate space, while the audience forms a distinct second layer in the background.

One area of interest is the application of proxemics to cinematography where the **shot composition** and **camera viewpoint** is optimized for visual weight [1]. In cinema, a shot is either a *long shot*, a *medium shot* or a *close-up* depending on whether it depicts "public proxemics", "personal proxemics" or "intimate proxemic", respectively. Similarly, the camera viewpoint is chosen based on the degree of occlusion allowed in the scene. To assure full visibility of every character in the scene, a *high-angle shot* is chosen whereas for intimate scenes and closeups, an *eye-level shot* or *low-angle shot* is more suitable.

From these attributes, we can see that each of the interactions specified in Figure 1 can be described as a combination of several of these factors. For example, "Group Interaction" in Figure 1(a) shows people within social distance in a single layer with a concave shape and is captured using high-angle, medium shot. In contrast, a "Family photo" in Figure 1(b) is an eye-level, closeup shot of a family within intimate distance. The taxonomy of attributes shown in Figure 3 are used to design features for Visual Proxemics classification, as discussed in Section 3.3.

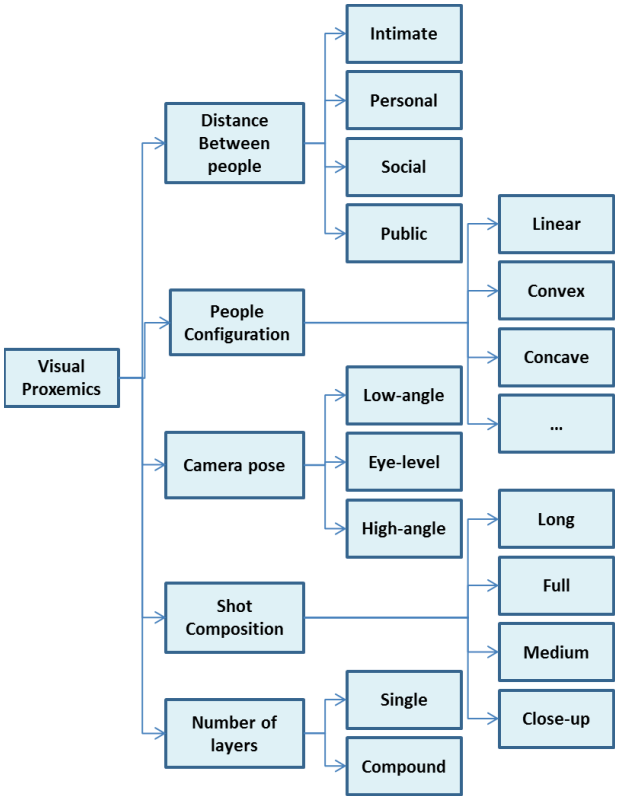


Figure 3: Taxonomy of attributes for Visual Proxemics.

Table 1: Statistics of our Visual Proxemes dataset based on NIST TRECVID corpus [2].

Proxeme type	**# examples/# dataset**	**% dataset**
Group photo	345 / 3814	9.0%
Group interaction	187 / 3814	4.9%
Couple and audience	99 / 3814	2.6%
Crowd	2448 / 3814	64.2%
Family photo	148 / 3814	3.8%
Speaker and audience	68 / 3814	1.8%
Undefined	519 / 3814	13.6%
High-angle	918 / 3814	24%
Eye-level	2722 / 3814	71%
Low-angle	174 / 3814	5%

3.2. Perspective Rectification Module

Given the 2D face locations in an image, the goal is to recover the camera height and the face positions in the X-Z plane relative to the camera center. These parameters are computed by using the camera model described in [14] and iterating between the following two steps - 1. Initializing the model with coarse

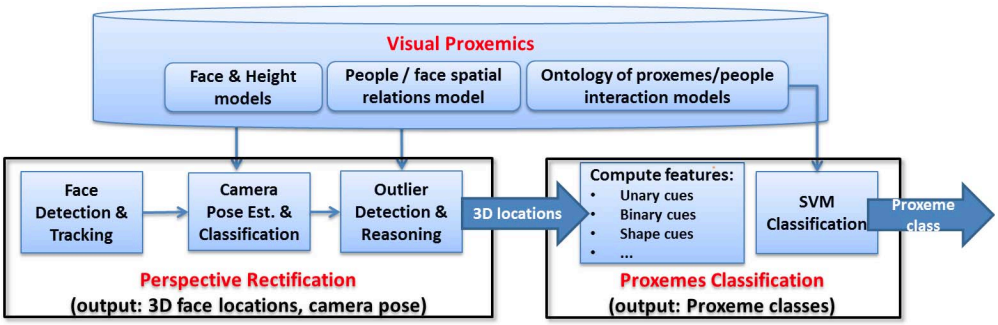


Figure 2: 3D Visual Proxemics Analysis (VPA3D) system diagram.

parameter estimates through a robust estimation technique. In addition to the parameters, we also detect outliers; face locations that do not fit the model hypothesis of uniform people heights and poses.This is described as the **outlier detection** step. 2. Refining the parameter estimates by 3D reasoning about position of outliers in relation to the inliers based on domain constraints that relate people's heights and poses. This is called the **outlier reasoning** step. The model alternates between estimating camera parameters and applying positional constraints until convergence is reached. We illustrate this approach in Figure 5. In the following sections, these two steps are described in detail.

3.2.1 RANSAC based Outlier Detection

This section describes an algorithm to estimate face depths, horizon line and camera height from 2D face locations in an image. Our model is based on the camera model described in [14]. The derivation is variously adapted from the presentations in [14, 10, 3]. We provide the derivation explicitly for the sake of completeness.

The coordinate transformation of a point using a typical pinhole camera model with uniform aspect ratio, zero skew and restricted camera rotation is given by,

$$\begin{pmatrix} u^i \\ v^i \\ 1 \end{pmatrix} = \frac{1}{z^w} \begin{pmatrix} f^w & 0 & u_c^w \\ 0 & f^w & v_c^w \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & cos\theta_x^w & -sin\theta_x^w & y_c^w \\ 0 & sin\theta_x^w & cos\theta_x^w & 0 \end{pmatrix} \begin{pmatrix} x^w \\ y^w \\ z^w \\ 1 \end{pmatrix}$$

where a (u^i, v^i) are its image coordinates, (x^w, y^w, z^w) are its 3D coordinates, and (u_c^w, v_c^w) are the coordinates of the camera center[2]. We assume that the camera is located at ($x_c^w = 0$,$z_c^w = 0$) and tilted slightly along x axis by θ_x^w. y_c^w is the camera height and f^w is the focal length.

At this stage some simplifying assumptions are made - (a) faces have constant heights, (b) faces rest on ground plane, which implies $y^w = 0$. The grounded position projects onto the bottom edge of the face bounding box in the image, $u^i = u_b^i, v^i = v_b^i$. (c) camera tilt is small, which implies $cos\theta_x \sim 1$ and $sin\theta_x \sim \theta_x \sim tan\theta_x \sim (v_c^w - v_0^i)/f$, where v_0^i is the height of the horizon line (also known as vanishing line) in image coordinates. By applying these approximations, we estimate z^w and x^w respectively,

$$z^w = \frac{f^w y_c^w}{(v_b^i - v_0^i)} \qquad x^w = \frac{y_c^w (u_b^i - u_c^w)}{(v_b^i - v_0^i)} \tag{1}$$

The estimated z^w is the 3D distance in depth from the camera center z_c^w and x^w is the horizontal distance from the camera center x_c^w. Using these (x^w,z^w) coordinates, we can undo the perspective projection of the 2D image and recover the *perspective rectified* face layout in the 3D coordinate system. Substituting the value of z^w into the equation for y^w and ignoring small terms we get,

$$y^w (v_b^i - v_0^i) = y_c^w (v^i - v_b^i) \tag{2}$$

This equation relates the world height of a face (y^w) to its image height ($v^i - v_b^i$) through its vertical position in the image (v_b^i) and through two unknowns - the camera height (y_c^w) and the horizon line (v_0^i). In general, given $N >= 2$ faces in an image, we have the following system of linear equations.

$$\begin{pmatrix} h_1 & h_w \\ ... & \\ h_N & h_w \end{pmatrix} \begin{pmatrix} y_c \\ v_0 \end{pmatrix} = \begin{pmatrix} h_w v_{b1} \\ ... \\ h_w v_{bN} \end{pmatrix}, \tag{3}$$

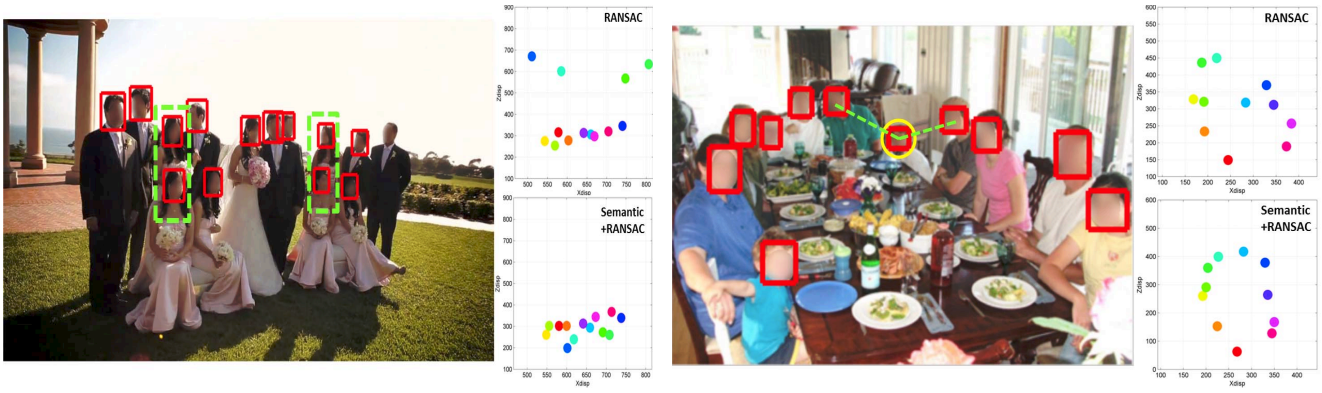
Thus, given an image with at least two detected faces, we can simultaneously solve for the two unknowns by minimizing the linear least squares error.

To get meaningful camera parameters, it is essential to filter out irregular observations that violate the model hypothesis. We use Random Sample Consensus (RANSAC) to reject these so-called outliers to get robust estimates. RANSAC is an iterative framework with two steps. First, a minimal sample set (2 face locations) is selected and model parameters $(\hat{z_w}, \hat{y_c}, \hat{v_0})$ are computed by least squares estimator (as explained above). Next, each instance of the observation set is checked for consistency with the estimated model. We estimate the face height in the image according to the model using $\hat{h_i} = h_w(v_b - \hat{v_0})/\hat{y_c}$ and compute the deviation from the observed height using $e_M^i = ||h_i - \hat{h_i}||$ to find the estimator error for that face. Outliers are instances whose summed errors over all the iterations exceed a pre-defined threshold.

3.2.2 Semantic Constraints based Outlier Reasoning

The linearized model is based on the hypothesis that all faces are (a) located on the same plane and (b) of the same size. However

[2]superscript w indicates 3D coordinates and i indicates image coordinates



(a) Visibility constraint

(b) Localized pose constraint

Figure 4: Circled faces depict outliers and the connected faces show the related inliers discovered through semantic constraints.

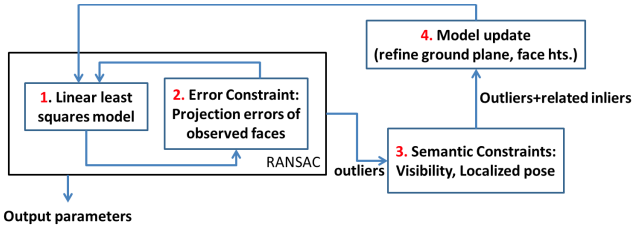


Figure 5: Flowchart of our Perspective Rectification module.

these assumptions do not always hold in practice. The faces that violate these assumptions are detected as outliers in the RANSAC step. Conventionally, outliers are treated as noisy observations and rejected from estimates. However, outlier faces may occur because of variations in face sizes and heights arising due to difference in age, pose (sitting versus standing) and physical planes of reference (ground level or on a platform). Hence, instead of eliminating them from consideration, we attempt to reason about them and restore them in our calculations. For doing this, we make use of semantics of Visual Proxemics to constrain the possible depth orderings of the outlier faces in the image. In particular, we consider two types of constraints - *visibility constraint* and *localized pose constraint*, as explained below.

Visibility Constraint

Consider the pose configuration in Figure 4(a). RANSAC estimates the sitting person's face to be an outlier because it violates the common ground plane assumption (assumption (b) in the linear model). However, we can easily see that for the sitting person is visible, she has to be in front of the standing person. We formulate this visibility constraint as follows - The only way for two faces to be visible at the same horizontal location is if the lower face is closer in depth than the face above it[3]. We formulate this constraint by the following inequality.

$$\delta(x_i - x_i^*)(y_i - y_i^*)(\hat{z_w} - \hat{z_w}^*) \leq 0, \tag{4}$$

[3] $(x = 0, y = 0)$ is upper left corner in image, z increases upwards.

where $\delta(a-b)$ is 1 when $a = b$ and $\hat{z_w}$ is the RANSAC estimate of depth. For each outlier in the image, we determine if it shares a visibility constraint with any of the inliers and maintain an index of all such pairs. Each such (outlier, inliers) pair is assumed to share a common ground plane (are standing/sitting on the same ground level). Based on this assumption the height estimates of the outliers are refined, as described in Section 3.2.3.

Localized pose Constraint

This constraint assumes that the people who are physically close to each other also share the same pose. Consider the group photo in Figure 4(b). RANSAC estimation (top plot) detects the child's face location as an outlier and incorrectly estimates its depth because of the height difference from the remaining members of the group. Now, if we assume that the inliers that are physically close to the outlier in the world also have a similar pose, then we can localize the ground plane level at the outlier based on the face locations of the neighboring inliers. This can help us fix the depth of the outlier without concerns about its vertical position in the image (as shown in the bottom plot).

Formally, let $N^{in}_{x^{out}_j}$ represent the inlier neighbors of outlier instance j along horizontal coordinates in the image. If the difference in the face size of the outlier to its inlier neighbors is within a threshold, then we can fix the depth of the outlier within the vicinity of the neighboring inliers. Formally, this constraint is represented as

$$(h^i_j - \sum h^i_{N^{in}_{x^{out}_j}} / N^{in}_{x^{out}_j}) < \epsilon^i_h \Rightarrow \tag{5}$$

$$(z^w_j - \sum z^w_{N^{in}_{x^{out}_j}} / N^{in}_{x^{out}_j}) < \epsilon^w_z \tag{6}$$

For each outlier in the image, we perform this constraint test to determine (outlier, inliers) pairs that satisfy the localized pose constraint. These are used to refine the height estimates of the outliers in the following section.

3.2.3 Model update

The height estimates of the outliers are refined using the semantically constrained set of inliers. Specifically, we make use of a piecewise constant ground plane assumption in the image to estimate the outlier heights in the world. By assuming that the outliers are located at the same level as the related inliers, the world height (h^w) of the outliers can calculated in proportion to the inliers. Let B_j^{out} is the body height of an outlier and $\hat{G_k^{in}}$ be the ground plane approximation for a neighboring inlier. The ground level is calculated by translating the vertical position of the face by a quantity proportional to the image height (we assume face size is 7 times the body size). The body height of the outlier is based on the average ground plane estimated from its inliers. The face height is then calculated as a fraction of the estimated body height.

$$\hat{G_k^{in}} = v_{bk}^i + c * h_k^i, \tag{7}$$

$$B_j^{out} = \frac{\sum_{k \in (in, N(j))} \hat{G_k^{in}}}{\sum_{k \in (in, N(j))}} - v_{bj}, \quad h_j^{out} = B_j^{out} / B_k^{in} \tag{8}$$

The new height ratios are inputs to the next round of RANSAC step that produce new estimates of face depths and camera heights. We perform this iteration 3-4 times in our model.

3.3. Visual Proxeme Description and Classification

To capture the spatial arrangement of people, we construct features based on the attribute taxonomy of Visual Proxemics presented in Section 3.1 (Figure 3). Specifically, our features are designed to reflect the information about the following attributes - 1) Distance, 2) Camera pose, 3) Shape, 4) Shot composition, and 5) Shape layers.

- **Shape cues**: The *Convex Hull* and the *Minimum Spanning Tree* (MST) of faces in X-Z provide cues about the extent and orientation of the shape. Specifically, we compute volume, width and height of the convex hull and the *eccentricity* of its envelope ellipse. We calculate the degree of MST (maximum and standard deviation) to determine the branching factor (BF); a high BF indicates a compact shape while low BF indicates a linear shape e.g., in a group photo.

- **Shape layers cues**: We find if people are arranged in a single group or in separate subgroups based on within and between layer distances and orientations. Specifically, we compute a) *affinity groups* by partitioning an affinity matrix. To generate the matrix, first find the pairwise distances between faces and normalize by the maximum distance for each face. Then, make the pairwise distances symmetric by averaging between each pair. Finally, the affinity matrix is partitioned to discover subgroups. b) *inter-face orientation*, for which we compute angles between face pairs along the MST with respect to X axis.

- **Shot composition cues**: We compute spatial distribution of people in the scene. We use the *number of points inside* the hull and the *ratio* between inside and outside points. Values $<< 1$ indicate high spread, e.g., as in a crowd. We also calculate - a) *Horizontal skew*: Using extremal face locations along X direction as anchors, the center and standard deviation along X axis. b) *Depth skew*: standard deviation of shape along Z axis, and c) *Centeredness*, which combines the deviations along X and Z axis.

- **Distance cues**: We measure the average Euclidean distance between face pairs in X-Z plane. Specifically, we consider two kinds of distances - a) *All-pairs distance*, between each pair of faces, normalized by the image diagonal. This indicates the average interpersonal distance in a group. b) *Nearest neighbor distance*, between faces along the MST. This captures the interaction distances.

- **Camera pose cues**: The camera height is quantized into three levels - *low-angle*, *eye-level* and *high-angle*. This cue captures the position of the cameraman with respect to the scene.

The raw features measure different types of statistics and thus lie on different scales. To fit the distribution of each feature within a common scale, we use a sigmoid function that converts feature values into probabilistic scores between zero and one. Some of these features are meaningful within a certain range of values. Therefore, shifting the sigmoid according to the threshold allows soft thresholding. Let σ be the threshold for feature x and c be the weight. Then, the following expression denotes the probabilistic value of the feature.

$$p(x) = \begin{cases} \frac{1}{1+e^{-c*(x-\sigma)}}, & \text{Threshold } x > \sigma \\ \frac{1}{1+e^{-c*(\sigma-x)}}, & \text{Threshold } x <= \sigma, \end{cases}$$

To compute an aggregate feature from all the faces in an image, we consider the mean and variance values of each feature and then fit the sigmoid function to re-adjust the values. The feature corresponding to an image is a concatenated vector of these probability scores.

4. Experiments and Results

We test our algorithms on the Visual Proxemics dataset presented in Table 1. Our experiments are directed towards evaluating the performance of 1) Camera parameter estimation from single view images of groups of people, namely, depth perception (Section 4.1) and camera height estimation (Section 4.2) and, 2) Visual Proxeme classification (Section 4.3). We mainly use average precision (A.P.) as the performance metric, unless specified otherwise. The A.P. is calculated using the standard 11-point interpolated method.

In the rectification algorithm, the standard face size is set to be 21 cm and the focal length is computed as 1.54 times the image height, based on the normal lens assumption [3]. The RANSAC step is run for 20 iterations and the overall estimation is run for 4 iterations. For proxeme classification, LibSVM [4] is used as the SVM solver and cross validation was used to determine the parameters.

4.1. Depth Perception

To evaluate the advantage of our overall rectification module, we compare the final results vis-a-vis the estimates we get before applying the semantically derived constraints i.e., with linear model and RANSAC based robust estimator. Additionally, as a baseline we consider the scale-ratios method in which Hoiem et al. [13] relate scale changes to an explicit 3D information.

In our test, we compare the estimated depths from algorithms with the depths perceived by a human annotator. In each image, each face is treated as an anchor and all other faces are color coded according to whether they are estimated to be ahead, behind or at the same depth as the face anchor. The human annotator then verifies the decisions of the algorithms and counts the number of errors per image. We report results on 60 test images with an average of 8.6 faces per image. The results are reported in Table 2.

Table 2: Comparison of depth ordering accuracy on 517 faces. The results from our complete framework outperforms the scale-ratios method [13] and our results from the intermediate steps of our algorithm.

Algorithm	#Incorr. alignments	%Incorr. per image
Scale-ratios [13]	1490	25.68
Step 1: Linear	1324	22.44
Step 2: RANSAC	836	13.93
Final: Semantic	**658**	**10.96**

The scale-ratios method systematically misconstrues depth estimates because it only uses face size and ignores the location in the image. The location is specially informative in high-angle and low-angle shots where size changes less with depth. The robust estimate with RANSAC performs better than scale-ratios. However, combining a robust model with proxemic semantics clearly outperforms other methods in its ability to handle wide variations in pose and heights of people.

4.2. Camera height estimation

Our model jointly computes 3D coordinates and camera height from faces in an image. As **baseline**, we consider the line-intersection method proposed in several earlier works, e.g., in [16]. In this method, given an image with upright humans, the line connecting the heads and the line connecting the feet intersects at a point on the horizon line. Then, a line is robustly fitted on the intersection points. We use this on the face bounding boxes and compute the horizon line for our baseline.

The images in our dataset are annotated by a human annotator with three camera views - low-angle (camera looking upwards at the scene), eye level, high-angle (camera looking downwards) (Table 1). The number of instances of eye-level views outnumbers the high or low-angle views. This is typical in Internet content where personal videos (associated with eye-level view) are more common than public gatherings e.g., parades and town hall meetings (associated with high or low-angle views). The results are presented in Table 3. This bias in the number of exemplars affects the overall precision of correct retrieval. Another parameter that affects the performance is the effectiveness of face detectors in high-angle/low-angle shots. We noticed a large drop in performance of face detection on small face sizes and non-frontal face shots, which are typical in images of large groups. Because of these reasons the eye-level shots are best detected by our system. However, our algorithm clearly outperforms the line intersection method.

4.3. Visual Proxeme Classification

The classes of visual proxemes that we consider are listed in Table 1. A group of human annotators mined through a large number of image keyframes of people and decided on this set of 6 commonly occurring visual proxemes. A different human annotator classified each image into one of the 6 visual proxemes or as unknown. Using the features described in Section 3.3, we build a 29-dimensional feature vector of probabilistic scores corresponding to each feature. Our method is labeled as **3DShape** in Table 4. As **baseline**, we consider the spatial features proposed in Gallagher et al. [10]. Their features are mainly based on 2D face locations and an additional cue for predicting face size. This cue can be shown to be derived from the linear model described in Section 3.2.1 by tying parameters to model the size information. We call this a *Semi3D* model. We consider the 2D image-based features from [10] to construct a 10-dimensional feature vector (**Image2D**) and then append this vector with additional 5 dimensions derived from the **Semi3D** model to generate a 18-dimensional feature vector. We apply a binary, one-versus-all SVM classifier with an RBF kernel on a $60\% - 40\%$ split of the dataset for training and testing, respectively. The average precisions from 5 rounds of random splits are reported in Table 4.

Table 3: Comparison of camera height estimation from our framework and from the vertical-object intersection method proposed in [16]

	Low-angle	Eye-level	High-angle
# instances	174	2722	918
A.P.(Baseline [16])	12.87%	30.23%	25.90%
A.P.(Our method)	**41.09%**	**83.68%**	**63.60%**

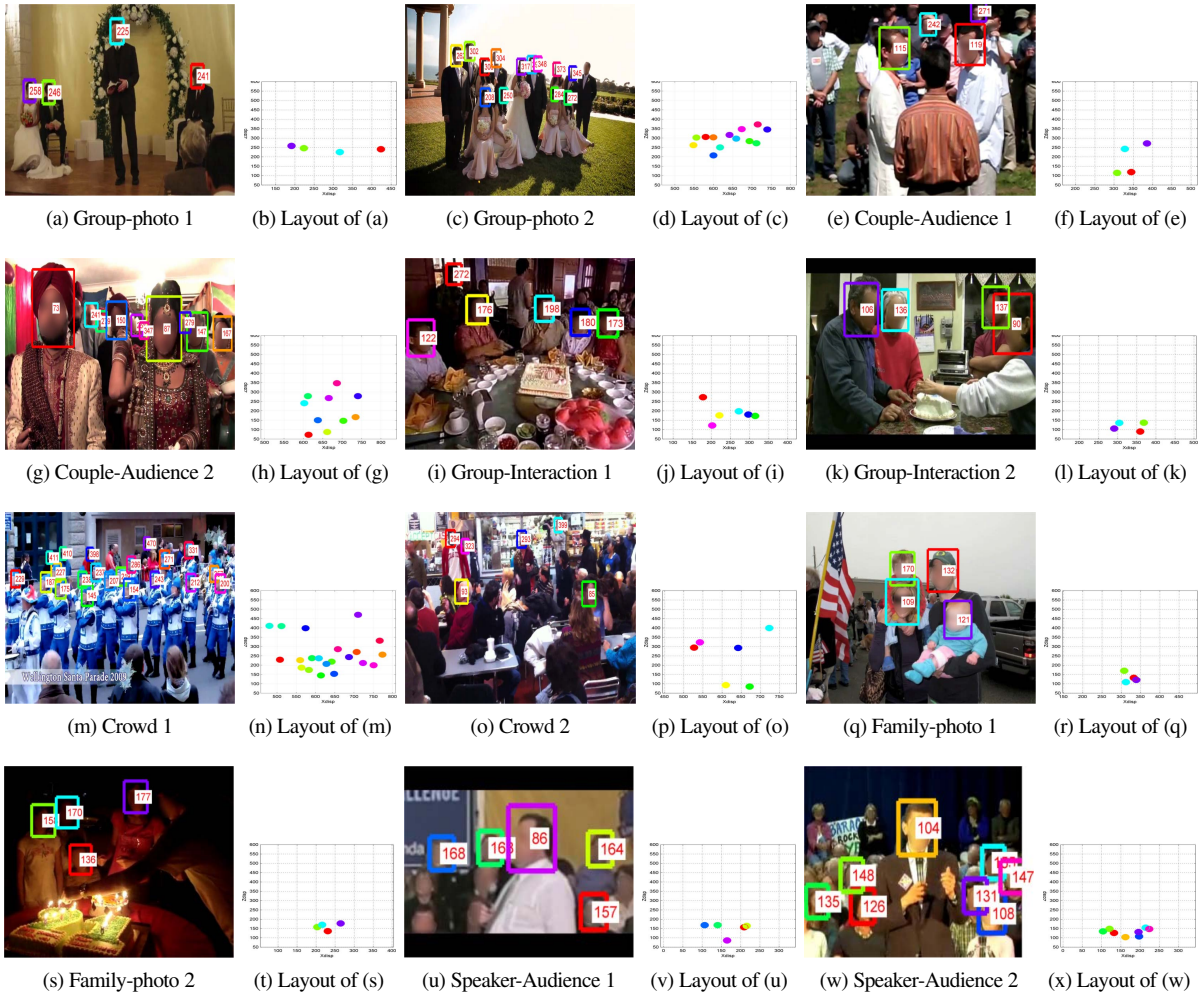
Table 4: Comparison of average precision of SVM based visual proxeme classification.

	Image2D	**Semi3D** [10]	**3DShape**
Group photo	56.4%	**69.6%**	67.4%
Couple+Audience	41.2%	31.6%	**58.7%**
Group interaction	46.8%	42.0%	**59.2%**
Crowd	81.1%	83.0%	**85.7%**
Family photo	49.1%	48.3%	**63.6%**
Speaker+Audience	28.1%	52.8%	**87.7%**

In general, our shape features computed from the 3D layout of faces (3DShape) performs best for all the classes, except one. It is interesting to note that the performance benefit is maximized in concepts that contain strong 3D cues. For example, Speaker+Audience, Couple+Audience, Group-Interaction are the top three concepts with maximum improvement vis-a-vis other methods. Our features best detect speaker+audience, which performs very poorly with other features. Figure 6 shows some of the true positives detected by the classifier and their corresponding layout in X-Z space.

5. Discussion

In this paper we present 3D Visual Proxemics Analysis, a framework that integrates Visual Proxemics with 3D arrangements of people to identify typical social interactions in Internet images. Our results demonstrate that this unified approach surpasses the state-of-the-art both in 3D estimation of people layout from detected faces as well as in classification of social interactions. We



(a) Group-photo 1 (b) Layout of (a) (c) Group-photo 2 (d) Layout of (c) (e) Couple-Audience 1 (f) Layout of (e)

(g) Couple-Audience 2 (h) Layout of (g) (i) Group-Interaction 1 (j) Layout of (i) (k) Group-Interaction 2 (l) Layout of (k)

(m) Crowd 1 (n) Layout of (m) (o) Crowd 2 (p) Layout of (o) (q) Family-photo 1 (r) Layout of (q)

(s) Family-photo 2 (t) Layout of (s) (u) Speaker-Audience 1 (v) Layout of (u) (w) Speaker-Audience 2 (x) Layout of (w)

Figure 6: Shows examples of true positives and the X-Z layouts detected by the SVM classifier.

believe that inclusion of semantics allowed us to estimate better 3D layout than the purely statistical approaches. A better 3D geometry, in turn, allowed us to define features informed by Proxemics that improved our semantic classification. In future, we hope to delve deeper into this synergistic approach by adding other objects and expanding our semantic vocabulary of Visual Proxemics.

References

[1] Wikipedia entry on proxemics.

[2] Nist trecvid multimedia event detection, June 2012.

[3] R. Baur, A. Efros, and M. Hebert. Statistics of 3d object locations in images. 2008.

[4] C. Chang and C. J. Lin. Libsvm : a library for support vector machines. In *ACM T-IST*, 2011.

[5] Y. Chen, W. Hsu, and H. Liao. Discovering informative social subgraphs and predicting pairwise relationships from group photos. In *ACM Multimedia*, 2012.

[6] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *CVPR*, 2011.

[7] R. Cipolla, T. Drummond, and D. Robertson. Camera calibration from vanishing points in images of architectural scenes. In *BMVC*, 1999.

[8] X. Cui, Q. Liu, M. Gao, and D. Metaxas. Abnormal detection using interaction energy potentials. In *CVPR*, 2011.

[9] A. Fathi, J. Hodgins, and J. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012.

[10] A. Gallagher and T. Chen. Understanding images of groups of people. In *CVPR*, pages 256–263, 2009.

[11] E. Hall. A system for the notation of proxemic behavior. *American Anthropologist*, 1963.

[12] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, 2005.

[13] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.

[14] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 80(1), 2008.

[15] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *CVPR*, 2010.

[16] F. Lv, T. Zhao, and R. Nevatia. Self-calibration of a camera from video of a walking human. In *IEEE PAMI*, 2006.

[17] M. Marin-Jimenez, A. Zisserman, and V. Ferrari. Heres looking at you, kid. detecting people looking at each other in videos. In *BMVC*, 2011.

[18] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009.

[19] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *CVPR*, 2012.