Understanding Images of Groups of People

Andrew C. Gallagher
Carnegie Mellon University
Pittsburgh, Pennsylvania
agallagh@cmu.edu

Tsuhan Chen
Carnegie Mellon University
Pittsburgh, Pennsylvania
tsuhan@ece.cornell.edu

Abstract

In many social settings, images of groups of people are captured. The structure of this group provides meaningful context for reasoning about individuals in the group, and about the structure of the scene as a whole. For example, men are more likely to stand on the edge of an image than women. Instead of treating each face independently from all others, we introduce contextual features that encapsulate the group structure locally (for each person in the group) and globally (the overall structure of the group). This "social context" allows us to accomplish a variety of tasks, such as such as demographic recognition, calculating scene and camera parameters, and even event recognition. We perform human studies to show this context aids recognition of demographic information in images of strangers.

1. Introduction

It is a common occurrence at social gatherings to capture a photo of a group of people. The subjects arrange themselves in the scene and the image is captured, as shown for example in Figure 1. Many factors (both social and physical) play a role in the positioning of people in a group shot. For example, physical attributes are considered, and physically taller people (often males) tend to stand in the back rows of the scene. Sometimes a person of honor (e.g. a grandparent) is placed closer to the center of the image as a result of social factors or norms. To best understand group images of people, the factors related to how people position themselves in a group must be understood and modeled.

We contend that computer vision algorithms benefit by considering *social context*, a context that describes people, their culture, and the social aspects of their interactions. In this paper, we describe contextual features from groups of people, one aspect of social context. There are several justifications for this approach. First, the topic of the spacing between people during their interactions has been thoroughly studied in the fields of anthropology [14] and social psychology [2]. A comfortable spacing between people depends on social relationship, social situation, gender and culture. This concept, called proxemics, is considered in architectural design [16, 26] and we suggest computer vision can benefit as well. In our work, we show experimental results that our contextual features from group images improves understanding. In addition, we show that human vision perception exploits similar contextual clues in interpreting people images.

Figure 1. Just as birds naturally space themselves on a wire (Upper Left), people position themselves in a group image. We extract contextual features that capture the structure of the group of people. The nearest face (Upper Right) and minimum spanning tree (Lower Left) both capture contextual information. Among several applications, we use this context to determine the gender of the persons in the image (Lower Right).

We propose contextual features that capture the structure of a group of people, and the position of individuals within the group. A traditional approach to this problem might be to detect faces and independently analyze each face by extracting features and performing classification. In our approach, we consider context provided by the global structure defined by the collection of people in the group. This allows us to perform or improve several tasks such as: identifying the demographics (ages and genders) of people in

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

the image, estimating the camera and scene parameters, and classifying the image into an event type.

2. Related Work

A large amount of research addresses understanding images of humans, addressing issues such as recognizing an individual, recognizing age and gender from facial appearance, and determining the structure of the human body. The vast majority of this work treats each face as an independent problem. However, there are some notable exceptions. In [4], names from captions are associated with faces from images or video in a mutually exclusive manner (each face can only be assigned one name). Similar constraints are employed in research devoted to solving the face recognition problem for consumer image collections. In [10, 19, 24], co-occurences between individuals in labeled images are considered for reasoning about the identities of groups of people (instead of one person at a time). However, the co-occurence does not consider any aspect of the spatial arrangement of the people in the image. In [23], people are matched between multiple images of the same person group, but only appearance features are used. Facial arrangement was considered in [1], but only as a way to measure the similarity between images.

Our use of contextual features from people images is motivated by the use of context for object detection and recognition. Hoiem *et al*. [15], and Torralba and Sinha [25] describe the context (in 3D and 2D, respectively) of a scene and the relationship between context and object detection. Researchers recognize that recognition performance is improved by learning reasonable object priors, encapsulating the idea that cars are on the road and cows stand on grass (not trees). Learning these co-occurence, relative co-locations, and scale models improves object recognition [11, 20, 21, 22]. These approaches are successful because the real world is highly structured, and objects are not randomly scattered throughout an image. Similarly, there is structure to the positions of people in a scene that can be modeled and used to aid our interpretation of the image.

Our contribution is a new approach for analyzing images of multiple people. We propose features that relate to the structure of a group of people and demonstrate that they contain useful information. The features provide social context that allows us to reason effectively in different problem domains, such as estimating person demographics, estimating parameters related to scene structure, and even categorizing the event in the image. In Section 3, we describe our image collection. In Section 4, we introduce contextual person features, and we detail their performance for classifying person demographics. We introduce the concept of a *face plane* and demonstrate its relationship to the scene structure and event semantics (Section 6). Finally, in Section 7 we describe experiments related to human perception based on cues related to social context.

	0-2	3-7	8-12	13-19	20-36	37-65	66+
Female	439	771	378	956	7767	3604	644
Male	515	824	494	736	7281	3213	609
Total	954	1595	872	1692	15048	6817	1253

Table 1. The distribution of the ages and genders of the 28231 people in our image collection.

3. Images and Labeling

We built a collection of people images from Flickr images. As Flickr does not explicitly allow searches based on the number of people in the image, we created search terms likely to yield images of multiple people. The following three searches were conducted:
"`wedding+bride+groom+portrait`"
"`group shot`" or "`group photo`" or "`group portrait`"
"`family portrait`"
A standard set of negative query terms were used to remove undesirable images. To prevent a single photographer's images from over-representation, a maximum of 100 images are returned for any given image capture day, and this search is repeated for 270 different days.

In each image, we labeled the gender and the age category for each person. As we are not studying face detection, we manually add missed faces, but 86% of the faces are automatically found. We labeled each face as being in one of seven age categories: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+, roughly corresponding to different life stages. In all, 5,080 images containing 28,231 faces are labeled with age and gender (see Table 1), making this what we believe is the largest dataset of its kind [9]. Many faces have low resolution. The median face has only 18.5 pixels between the eye centers, and 25% of the faces have under 12.5 pixels.

As is expected with Flickr images, there is a great deal of variety. Some images have people are sitting, laying, or standing on elevated surfaces. People often have dark glasses, face occlusions, or unusual facial expressions. Is there useful information in the structure and arrangement of people in the image? The rest of the paper is devoted to answering this question to the affirmative.

4. Contextual Features from People Images

A face detector and an Active Shape Model [6] are used to detect faces and locate the left and right eye positions. The position $\mathbf{p} = \begin{bmatrix} x_i & y_i \end{bmatrix}^T$ of a face f is the two dimensional centroid of the left and right eye center positions $\mathbf{l} = \begin{bmatrix} x_l & y_l \end{bmatrix}^T$ and $\mathbf{r} = \begin{bmatrix} x_r & y_r \end{bmatrix}^T$:

$$\mathbf{p} = \frac{1}{2}\mathbf{l} + \frac{1}{2}\mathbf{r} \tag{1}$$

The distance between the two eye center positions for the face is the size $e = ||\mathbf{l} - \mathbf{r}||$ of the face. To capture the structure of the people image, and allow the structure of the group to represent context for each face, we compute the following features and represent each face $\mathbf{f}_x$ as a 12-dimensional contextual feature vector:

Absolute Position: The absolute position of each face $\mathbf{p}$, normalized by the image width and height, represents two dimensions. A third dimension in this category is the angle of the face relative to horizontal.

Relative Features: The centroid of all the faces in an image is found. Then, the relative position of a particular face is the position of the face to the centroid, normalized to the mean face size:

$$\mathbf{r} = \frac{\mathbf{p} - \mathbf{p}_\mu}{e_\mu} \tag{2}$$

where $\mathbf{r}$ is the relative position of the face, $\mathbf{p}_\mu$ is the centroid of all faces in the image, and e_μ is the mean size of all faces from the image. The third dimension in this category is the ratio of the face size to the mean face size:

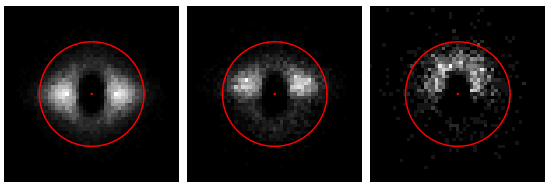
$$e_r = \frac{e}{e_\mu} \tag{3}$$

When three or more faces are found in the image, a linear model is fit to the image to model the face size as a function of y-axis position in the image. This is described in more detail in Section 5.2. Using (9), the predicted size of the face compared with the actual face size is the last feature:

$$e_p = \frac{e}{\alpha_1 y_i + \alpha_2} \tag{4}$$

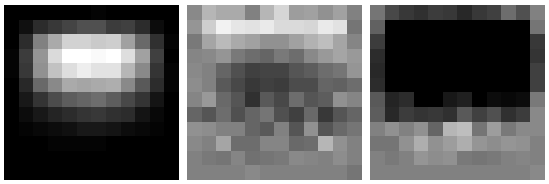
Minimal Spanning Tree: A complete graph $G = (V, E)$ is constructed where each face $\mathbf{f}_n$ is represented by a vertex $v_n \in V$, and each edge $(v_n, v_m) \in E$ connects vertices v_n and v_m. Each edge has a corresponding weight $w(v_n, v_m)$ equal to the Euclidean distance between the face positions $\mathbf{p}_n$ and $\mathbf{p}_m$. The minimal spanning tree of the graph $MST(G)$ is found using Prim's algorithm. The minimal spanning tree reveals the structure of the people image; if people are arranged linearly, the minimal spanning tree $MST(G)$ contains no vertices of degree three or greater. For each face $\mathbf{f}_n$, the degree of the vertex v_n is a feature $\deg(v_n)$. An example tree is shown in Figure 1.

Nearest Neighbor: The K nearest neighbors, based again on Euclidean distance between face positions $\mathbf{p}$ are found. As we will see, the relative juxtaposition of neighboring faces reveals information about the social relationship between them. Using the nearest neighbor face, the relative position, size, and in-plane face tilt angle are calculated, for a total of four dimensions.



(a) All (b) Female-Male (c) Baby-Other

Figure 2. The position of the nearest face to a given face depends on the social relationship between the pair. (a) The relative position of two nearest neighbors, where the red dot represents the first face, and lighter areas are more likely positions of the nearest neighbor. The red circle represents a radius of 1.5 feet (457mm). (b) When nearest neighbors are male and female, the male tends to be above and to the side of the female (represented by the red dot). (b) The position of the nearest neighbor to a baby. The baby face tends to be spatially beneath the neighbor, and incidentally, the nearest neighbor to a baby is a female with probability 63%.



(a) $P(\mathbf{p})$ (b) $P(f_g = \mathtt{m}|\mathbf{p})$ (c) $P(f_a < 8|\mathbf{p})$

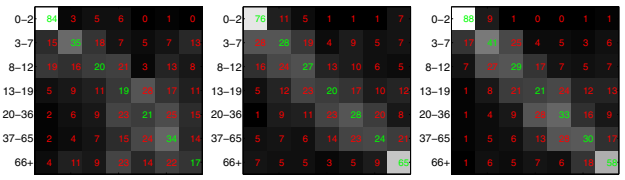
Figure 3. The absolute position of a face in the image provides clues about age and gender. Each of the three images represent a normalized image. (a) The density of all 28231 faces in the collection. (b) $P(f_g = \mathtt{male}|\mathbf{p})$. A face near the image edge or top is likely to be male. (c) $P(f_a < 8|\mathbf{p})$. A face near the bottom is likely to be a child.

The feature vector $\mathbf{f}_x$ captures both the pairwise relationships between faces and a sense of of the person's position relative to the global structure of all people in the image.

4.1. Evidence of Social Context

It is evident the contextual feature $\mathbf{f}_x$ captures information related to demographics. Figure 2 shows the spatial distributions between nearest neighbors. The relative position is dependent on gender (b) and age (c). Using the fact that the distance between human adult eye centers is 61$\pm$3 mm [8], the mean distance between a person and her nearest neighbor is 306 mm. This is smaller than the 18-inch radius "personal space" of [2], but perhaps subjects suspend their need for space for the sake of capturing an image.

Figure 3 shows maps of $P(f_a|\mathbf{p})$ and $P(f_g|\mathbf{p})$, the probability that a face has a particular gender or age given absolute position. Intuitively, physically taller men are more likely to stand in the group's back row and appear closer to the image top. Regarding the degree $\deg(v_n)$ of a face in $MST(G)$, females tend to be more centrally located in



(a) Context (b) Appearance (c) Both

Figure 4. The structure of people in an image provides context for estimating age. Show are the confusion matrices for classifying age using (a) context alone (no face appearance), (b) content (facial appearance) alone, (c) both context and facial appearance. Context improves over content alone.

	Gender	Age	
Random Baseline	50.0%	14.3%	38.8%
Absolute Position	62.5%	25.7%	56.3%
Relative Position	66.8%	28.5%	60.5%
Min. Spanning Tree	55.3%	21.4%	47.2%
Nearest Neighbor	64.3%	26.7%	56.3%
Combined $\mathbf{f}_x$	**66.9%**	**32.9%**	**64.4%**

Table 2. Predicting age and gender from context features $\mathbf{f}_x$ alone. The first age column is the accuracy for an exact match, and the second allows an error of one age category (e.g. a 3-7 year old classified as 8-12).

a group, and consequently have a higher mean degree in $MST(G)$. For faces with $\deg(v_n) > 2$ the probability the face is female is 62.5%.

4.2. Demographics from Context and Content

The interesting research question we address is this: How much does the structure of the people in images tell us about the people? We estimate demographic information about a person using $\mathbf{f}_x$. The goal is to estimate each face's age f_a and gender f_g. We show that age and gender can be predicted with accuracy significantly greater than random by considering only the context provided by $\mathbf{f}_x$ and no appearance features. In addition, the context has utility for combining with existing appearance-based age and gender discrimination algorithms.

4.2.1 Classifying Age and Gender with Context

Each face in the person image is described with a contextual feature vector $\mathbf{f}_x$ that captures local pairwise information (from the nearest neighbor) and global position. We trained classifiers for discriminating between age and gender. In each case, we use a Gaussian Maximum Likelihood (GML) classifier to learn $P(f_a|\mathbf{f}_x)$ and $P(f_g|\mathbf{f}_x)$. The distribution of each class (7 classes for age, 2 for gender) is learned by fitting a multi-variate Gaussian to the distributions $P(\mathbf{f}_x|f_a)$ and $P(\mathbf{f}_x|f_g)$. Other classifiers (Adaboost, decision forests, SVM) yield similar results on this problem, but GML has the advantage that the posterior is easy to directly estimate.

The age classifier is trained from a random selection of 3500 faces, selected such that each age category has an equal number of samples. Testing is performed on an independent (also uniformly distributed) set of 1050 faces. Faces for test images are selected to achieve roughly an even distribution over the number of people in the image. The prior for gender is roughly even in our collection, so we use a larger training set of 23218 images and test on 1881 faces.

For classifying age, our contextual features have an accuracy more than double random chance (14.3%), and gender is correctly classified about two-thirds of the time. Again, we emphasize that no appearance features are considered. Table 2 shows the performance of our classifiers for the different components of the contextual person feature $\mathbf{f}_x$. The strongest single component is Relative Position, but the inclusion of all features is the best. Babies are recognized with good accuracy, mainly because their faces are smaller and positioned lower than others in the image.

4.2.2 Combining Context with Content

We trained appearance-based age and gender classifiers. These content-based classifiers provide probability estimates $P(f_g|\mathbf{f}_a)$ and $P(f_a|\mathbf{f}_a)$ that the face has a particular gender and age category, given the visual appearance $\mathbf{f}_a$. Our gender and age classifiers were motivated by the works of [12, 13] where a low dimension manifold for the age data. Using cropped and scaled faces (61×49 pixels, with the scaling so the eye centers are 24 pixels apart) from the age training set, two linear projections ($\mathbf{W}_a$ for age and $\mathbf{W}_g$ for gender) are learned. Each column of $\mathbf{W}_a$ is a vector learned by finding the projection that maximizes the ratio of interclass to intraclass variation (by linear discriminate analysis) for a pair of age categories, resulting in 21 columns for $\mathbf{W}_a$. A similar approach is used to learn a linear subspace for gender $\mathbf{W}_g$. Instead of learning a single vector from two gender classes, a set of seven projections is learned by learning a single projection that maximizes gender separability for each age range.

The distance d_{ij} between two faces is measured as:

$$d_{ij} = (\mathbf{f}_i - \mathbf{f}_j)\mathbf{W}\mathbf{W}^T(\mathbf{f}_i - \mathbf{f}_j)^T \quad (5)$$

For classification for both age and gender, the nearest N training samples (we use $N = 25$) are found in the space defined by $\mathbf{W}_a$ for age or $\mathbf{W}_g$ for gender. The class labels of the neighbors are used to estimate $P(f_a|\mathbf{f}_a)$ and $P(f_g|\mathbf{f}_g)$ by MLE counts. One benefit to this approach is that a common algorithm and training set are used for both tasks, only the class labels and pairing for learning discriminative projections are modified.

The performance of both classifiers seems reasonable given the difficulty of this collection. The gender classifier

	Gender	Age	
Context $\mathbf{f}_x$	66.9%	32.9%	64.4%
Appearance $\mathbf{f}_a$	69.6%	38.3%	71.3%
Combined $\mathbf{f}_x, \mathbf{f}_a$	**74.1%**	**42.9%**	**78.1%**

Table 3. In images of multiple people, age and gender estimates are improved by considering both appearance and the social context provided by our features. The first age column is exact age category accuracy; the second allows errors of one age category.

	Gender	Age	
Context $\mathbf{f}_x$	65.1%	27.5%	63.5%
Appearance $\mathbf{f}_a$	67.4%	30.2%	65.9%
Combined $\mathbf{f}_x, \mathbf{f}_a$	**73.4%**	**36.5%**	**74.6%**

Table 4. For smaller faces $\leq$18 pixels between eye centers, classification suffers. However, the gain provided by combine context with content increases.

is correct about 70% of the time. This is lower than others [3], but our collection contains a substantial number of children, small faces and difficult expressions. For people aged 20-65, the gender classification is correct 75%, but for ages between 0-19, performance is a poorer 60%, as facial gender differences are not as apparent. For age, the classifier is correct 38% of the time, and if a one-category error is allowed, the performance is 71%. These classifiers may not be state-of-the-art, but are sufficient to illustrate our approach. We are interested in the *benefit* that can be achieved by modeling the social context.

Using the Naïve Bayes assumption, the final estimate for the class (for example, gender f_g) given all available features (both content $\mathbf{f}_a$ and context $\mathbf{f}_x$) is:

$$P(f_g|\mathbf{f}_a, \mathbf{f}_x) = P(f_g|\mathbf{f}_a)P(f_g|\mathbf{f}_x) \tag{6}$$

Table 3 shows that both gender and age estimates are improved by incorporating both content (appearance) and context (the structure of the person image). Gender recognition improves by 4.5% by considering person context. Exact age category recognition improves by 4.6%, and when the adjacent age category is also considered correct, the improvement is 6.8%. Figure 5 shows the results of gender classification in image form, with discussion. Accuracy suffers on smaller faces, but the benefit provided by context increases, as shown in Table 4. For example, context now improves gender accuracy by 6%. This corroborates [20] in that the importance of context increases as resolution decreases.

5. Scene Geometry and Semantics from Faces

The position of people in an image provides clues about the geometry of the scene. As shown in [18], camera calibration can be achieved from a video of a walking human, under some reasonable assumptions (that the person walks on the ground plane and head and feet are visible). By making broader assumptions, we can model the geometry of the scene from a group of face images. First, we assume faces approximately define a plane we call the *face plane*, a world plane that passes through the heads (i.e. the centroids of the eye centers) of the people in the person image. Second, we assume that head sizes are roughly similar. Third, we assume the camera has no roll with respect to the face plane. This ensures the face plane horizon is level. In typical group shots, this is approximately accomplished when the photographer adjusts the camera to capture the group.

Criminisi *et al*. [7] and Hoiem *et al*. [15] describe the measurement of objects rooted on the ground plane. In contrast, the face plane is not necessarily parallel to the ground, and many times people are either sitting or are not even on the ground plane at all. However, since the true face sizes of people are relatively similar, we can compute the face horizon, the vanishing line associated with the face plane.

5.1. Modeling the Face Plane

From the set of faces in the image, we compute the face horizon and the camera height (the distance from the camera to the face plane measured along the face plane normal), not the height of the camera from the ground. Substituting the face plane for the ground plane in Hoiem *et al*. [15], we have:

$$E_i = \frac{e_i Y_c}{y_i - y_o} \tag{7}$$

where E_i is the face inter-eye distance in the world (61 mm for the average adult), e_i is the face inter-eye distance in the image, Y_c is the camera height, y_i is the y-coordinate of the face center $\mathbf{p}$, and y_o is the y-coordinate of the face horizon.

Each of the N face instances in the image provides one equation. The face horizon y_o and camera height Y_c are solved using least squares by linearizing (7) and writing in matrix form:

$$\begin{bmatrix} E_{i1} & e_{i1} \\ E_{i2} & e_{i2} \\ \dots & \dots \\ E_{iN} & e_{iN} \end{bmatrix} \begin{bmatrix} y_o \\ Y_c \end{bmatrix} = \begin{bmatrix} y_{i1}E_{i1} \\ y_{i2}E_{i2} \\ \dots \\ y_{iN}E_{iN} \end{bmatrix} \tag{8}$$

Reasonable face vanishing lines and camera height estimates are produced, although it should be noted that the camera focal length is not in general recovered. A degenerate case occurs when the face plane and image planes are parallel (e.g. a group shot of standing people of different heights), the face vanishing line is at infinity, and the camera height (i.e. in this case, the distance from the camera to the group) cannot be recovered.

To quantify the performance of the camera geometry estimates, we consider a set of 18 images where the face vanishing plane and ground plane are parallel and therefore share a common vanishing line, the horizon. The horizon is

(a) Context (b) Appearance (c) Both (d) Context (e) Appearance (f) Both

(g) Context (h) Appearance (i) Both (j) Context (k) Appearance (l) Both

(m) Context (n) Appearance (o) Both (p) Context (q) Appearance (r) Both

Figure 5. Gender classification improves using context and appearance. The solid circle indicates the gender guess (pink for female, blue for male), and a dashed red line shows incorrect guesses. For the first four images (a)-(l), context helps correct mistakes made by the appearance classifier. The mislabeled men in (b) are taller than their neighbors, so context corrects their gender in (c), despite the fact that context has mistakes of its own (a). Similar effects can be seen in (d)-(l). The final two images (m)-(r) shows images where adding context degrades the result. In (p), context causes an incorrect gender estimate because the woman in on the edge and taller than neighbors even though the appearance classifier was correct (o). In (p)-(r), the people are at many different and apparently random depths, breaking the social relationships that are learned from training data. Best viewed in electronic version.

Figure 6. Horizon estimates from faces for images where the face and ground planes are approximately parallel. The solid green line shows the horizon estimate from the group of faces according to (8), and the dashed blue line shows the manually derived horizon (truth). The poor accuracy on the last image results from the four standing people, which violate the face plane assumption.

manually identified by finding the intersection in image coordinates of two lines parallel to the ground and each other (e.g. the edges of a dinner table). Figure 6 shows the estimated and ground truth horizons for several images, and the accuracy is reported in Table 5. Using the group shot face geometry achieves a median horizon estimate of 4.6%, improving from an error of 17.7% when the horizon is assumed to pass through the image center, or 9.5% when the horizon estimate is the mean position of all other labeled images. We experimented with RANSAC to eliminate difficult faces from consideration, but it made little difference in practice. We considered using the age and gender specific estimates for inter-eye distance values E_i, but this also resulted in a negligible gain in accuracy (<0.01%).

	Mean	Median
Center Prior	19.8%	17.7%
Mean Horizon Prior	9.6%	9.5%
Face Horizon	**6.3%**	**4.6%**

Table 5. The geometry of faces in group shots are used to accurately estimate the horizon. Mean and median absolute error (as percentage of image height) is shown for horizon estimates.

5.2. Event Recognition from Structure

Interestingly enough, the geometrical analysis of a group photo also represents context for semantic understanding. When a group shot is captured, the arrangement of people in the scene is related to the social aspect of the group. When a group is dining together, the face plane is roughly parallel

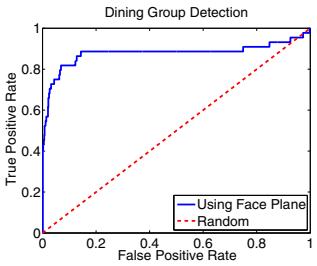


Figure 8. The face plane encapsulates semantically relevant information. The solid blue curve shows the detection of group dining images using a single feature related to the face plane. The red dashed curve shows expected random performance.

to the ground. In most other circumstances, a group photo contains a mixture of sitting and standing people at a nearly uniform distance from the camera, so the face plane is closer to orthogonal to the ground plane. An analysis of the face plane is useful for identifying the group structure and yields about the group activities.

We compute the value of $\frac{de_i}{dy_i}$, the derivative of face size with respect to position in the image. We use least squares to learn parameters α_1 and α_2 to model the face size as a function of position in the image according to:

$$e_i = \alpha_1 y_i + \alpha_2 \tag{9}$$

and then $\frac{de_i}{dy_i} = \alpha_1$. The model from (8) could also be used to estimate the size of a face in the face plane, but its objective function minimizes a quantity related to the camera and scene geometry and does not guarantee that the estimated face sizes in the image are optimal.

Figure 7 shows the ten images from the group photo collection with the most negative values of $\frac{de_i}{dy_i}$. Clearly, the structure of the face plane has semantic meaning. We perform an experiment to quantify this observation. Among the 826 "group photo" images with 5 or more people from the image collection, 44 are dining images. Using the single feature of $\frac{de_i}{dy_i}$, the group dining detection accuracy is shown in Figure 8. The good performance is somewhat remarkable given that dining images are recognized without explicitly looking for tables, plates, or any other features but facial arrangement. We find 61% of the dining images with a precision of 61%. This performance at least rivals that of [5] at detecting eating images (54%), even though they consider visual words and geographic location. This is a powerful demonstration that the structure in a people image provides important context for scene understanding.

6. Human Understanding

In the past, human accuracy for the age recognition task has been measured [12, 17], although the effect of context from other people in images on human performance has not been quantified. An experiment was designed to determine the role of context in the human interpretation of faces in group shots. Image content is progressively revealed in three stages as shown in Figure 9. In each stage, the subject must guess the age (in years) and gender of a face from a group photo. In the first stage, only the face (the same size as used by our appearance classifiers) of one individual is shown. Next, all faces in the image are revealed, and finally the entire image is shown. A subject enters age and gender estimates for all faces within a stage before progressing to the next stage.

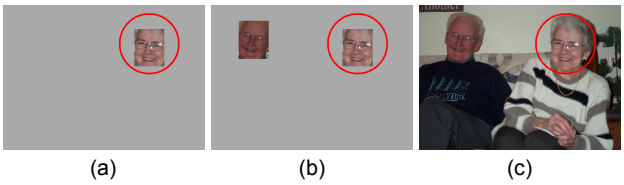


Figure 9. An example of the images shown to subjects in our human study. The subject estimates age and gender based on (a) the face alone, (b) all the faces in the image, or (c) the entire image. Human estimates of age and gender improve when additional context is available.

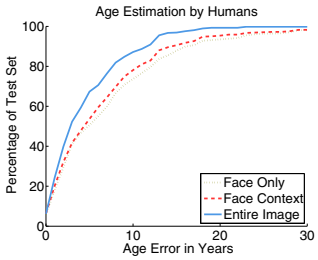


Figure 10. The effect of context on age estimation by humans. The curves show the percent of age estimates that are within a certain number of years in age error.

The 45 images for the experiment come from a dataset of personal image collections where the identity and birthdate of each person is known. True ages range from 2 to 83 years. A total of 13 subjects estimated age and gender for each of the 45 faces for each of the 3 stages, for a total of 1755 evaluations for age and gender.

The results are shown in Figure 10 and described in Table 6. Age prediction error is reduced as additional context is provided. Out of the 13 subjects, only 1 did not show an age error improvement from (a) to (b). Similarly, for the 45 face images, 33 show a reduction in age error from (a) to (b). Neither of these results could occur by chance with probability greater than 0.1%. As one might expect, estimating of a child's age can be achieved with better accuracy, but estimating the gender of a child is difficult from the face alone.

We draw several conclusions. First, human perception

Figure 7. Sorting images with respect to $\frac{de_i}{dy_i}$. The ten images with the most negative values of $\frac{de_i}{dy_i}$. These images tend to be ones where the face and ground planes are parallel, and often semantically correspond to group dining images (only the first, with a strong linear face structure, is a non-dining images).

	(a)	(b)	(c)
Mean Age Error	7.7	6.9	4.9
Children (< 13) Age Error	5.1	4.6	1.9
Adult (> 12) Age Error	8.1	7.3	5.5
Gender Error	6.2%	6.2%	1.0%
Children (< 13) Gender Error	15.4%	17.6%	0%
Adult (> 12) Gender Error	4.5%	4.0%	1.2%

Table 6. Human estimates of age and gender are more accurate with increasing amounts of context, from (a) face alone, to (b) all faces in the image, to (c) the entire image.

of faces benefits from considering social context. By simply revealing the faces of other people in the image, the subjects' age estimates improved, despite the fact that the additional viewable pixels were not on the person of interest. Finally, the experiment shows that the best estimates are achieved when the subject views the entire image and considers all the information to make demographic estimates.

7. Conclusion

In this paper we introduce contextual features for capturing the structure of people images. Instead of treating each face independently from all others, we extract features that encapsulate the structure of the group. Our features are motivated from research in several fields. In the social sciences, there is a long history of considering the spatial interactions between people. We provide evidence that our features provide useful social context for a diverse set of computer vision tasks. Specifically, we demonstrate gender and age classification, scene structure analysis, and event type classification to detect dining images. Finally, we show that even human understanding of people images benefits from the context of knowing who else in the image.

In a broader scope, we suggest that the social context in which an image is captured is relevant for its interpretation. We feel this is a rich area for researchers, and we provide our image collection to other interested researchers [9].

References

[1] M. Abdel-Mottaleb and L. Chen. Content-based photo album management using faces' arrangement. In *Proc. ICME*, 2004.

[2] I. Altman. *The Environment and Social Behavior: Privacy, Personal Space, Territory, Crowding*. Wadsworth Publishing Company, 1975.

[3] S. Baluja and H. Rowley. Boosting sex identification performance. In *IJCV*, 2007.

[4] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *Proc. CVPR*, 2004.

[5] L. Cao, J. Luo, H. Kautz, and T. Huang. Annotating collections of photos using hierarchical event and scene models. In *Proc. CVPR*, 2008.

[6] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *CVIU*, 1995.

[7] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *International Journal of Computer Vision*, 40:2000, 1999.

[8] L. Farkas. *Anthropometric facial proportions in medicine*. Raven Press, New York, 1994.

[9] A. Gallagher and T. Chen. The images of groups dataset. http://amp.ece.cmu.edu/people/andy/imagesOfGroups.html, 2009.

[10] A. C. Gallagher and T. Chen. Using group prior to identify people in consumer images. In *Proc. CVPR SLAM*, 2007.

[11] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location, and appearance. In *Proc. CVPR*, 2008.

[12] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In *ACM MULTIMEDIA*, 2006.

[13] G. Guo, Y. Fu, C. Dyer, and T. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. In *IEEE Trans. on Image Proc.*, 2008.

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

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

[16] W. Ju, B. Lee, and S. Klemmer. Range: Exploring proxemics in collaborative whiteboard interaction. In *Proc. CHI*, 2007.

[17] A. Lanitis, C. Dragonova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Trans. on Systems, Man and Cybernetics*, 2004.

[18] M.-F. Lv, M.-T. Zhao, and F.-R. Nevatia. Camera calibration from video of a walking human. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1513–1518, 2006.

[19] M. Naaman, R. Yeh, H. Garcia-Molina, and A. Paepcke. Leveraging context to resolve identity in photo albums. In *Proc. JCDL*, 2005.

[20] D. Parikh, L. Zitnick, and T. Chen. From appearance to context-based recognition: Dense labeling in small images. In *Proc. CVPR*, 2008.

[21] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proc. ECCV*, 2006.

[22] A. Singhal, J. Luo, and W. Zhu. Probabilistic spatial context models for scene content understanding. In *Proc. CVPR*, 2003.

[23] J. Sivic, C. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *Proc. BMVC*, 2006.

[24] Z. Stone, T. Zickler, and T. Darrell. Autotagging facebook: Social network context improves photo annotation. In *Proc. CVPR*, 2008.

[25] A. Torralba and P. Sinha. Statistical context priming for object detection. In *Proc. ICCV*, 2001.

[26] L. Veatch. Toward the environmental design of library buildings. *Library Trends*, 1987.