Using Context to Recognize People in Consumer Images

Andrew C. Gallagher and Tsuhan Chen

Abstract

Recognizing people in images is one of the foremost challenges in computer vision. It is important to remember that consumer photography has a highly social aspect. The photographer captures images not in a random fashion, but rather to remember or document meaningful events in her life. Understanding images of people necessitates that the context of each person in an image is considered. Context includes information related to the image of the scene surrounding the person, camera context such as location and image capture time, and the social context that describes the interactions between people.

The goal of this paper is to provide the computer with the same intuition that humans would use for analyzing images of people. Fortunately, rather than relying on a lifetime of experience, context can often be modeled with large amounts of publicly available data. Probabilistic graph models and machine learning are used to model the relationship between people and context in a principled manner.

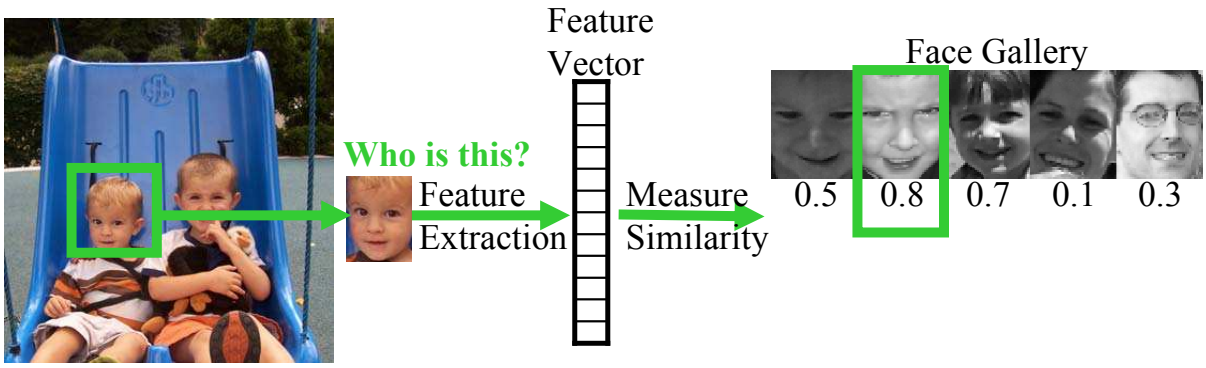


Figure 1: In traditional face recognition, identification is performed considering only face pixels. The face of interest is cropped, and features are extracted and compared to each face in a face gallery for identification. Using only local features makes identification difficult. The goal of this paper is to describe recent work for merging context with person recognition.

1 Introduction

Face recognition is one of the most important, yet difficult tasks in computer vision. Current methods generally focus on measuring the similarity between face images, as illustrated in Figure 1. Intuitively, recognition of the young boy would be aided by knowing the identity of his playmate, the location of the photo, the image capture date, the fact that he is sitting on a playground swing, and any other available contextual information. In contrast to a face-only approach, the human visual system brings with it a wealth of rich contextual information to recognize people in images, and faces are but one component of recognition for humans.

Figure 2 illustrates the limitations of using only facial features for recognizing people. When six faces (cropped and scaled in the same fashion as images from the PIE [28] database often are) from an image collection are shown, it is difficult to determine how many different individuals are present. Even if it is known that there are only three different individuals, the problem is not much easier. In fact, the three are sisters of similar age. When the faces are shown in context with their clothing, it becomes almost trivial to recognize which images are of the same person. To quantify the role clothing plays when humans recognize people, the following experiment was performed: 7 subjects were given a page showing 54 labeled faces of 10 individuals from the image collection and asked to identify a set of faces from the same collection. The experiment was repeated using images that included a portion of the clothing (as shown in Figure 2).

Figure 2: It is extremely difficult even for humans to determine how many different individuals are shown and which images are of the same individuals from only the faces (top). However, when the faces are embedded in the context of clothing, it is much easier to distinguish the three individuals (bottom).

The average correct recognition rate (on this admittedly difficult family album) jumped from 58% when only faces were used, to 88% when faces and clothing were visible. This experiment demonstrates the potential improvement for understanding images of people using context.

The goal of this paper is to model the relationship between context and person recognition in consumer images. The contextual information that a human uses to recognize people can often be modeled by fusing inference over image data, metadata (or "data about data") and large amounts of statistical data which model human interactions and social context. Probabilistic models and machine learning are used to integrate context into the interpretation of people in images. Applications for these ideas include:

- Recognizing people from descriptions, names, or a small number of labeled examples.
- Finding plausible names and demographic data for any person in an image collection.
- Understanding the relationships between people (friend vs. family vs. acquaintance).

We define context (Section 2), describe the related work (Section 3), briefly summarize some successful examples of considering context for person recognition (Section 4), and outline remaining work (Section 5).

2 What is Context?

Context is broadly defined as information relevant to something under consideration. Context can include information from other (i.e. non-face) regions of the image, information related to the capture of the image, or the social context of the interactions between people. Table 1 shows examples of context that are considered in our research.

Pixel context such as distinctive clothing or glasses can be useful for recognizing people in images. Further, because people tend to appear in images with friends and family, the identities of other people in an image aid our recognition of a person of interest. Even the position of a person in the image is important (for example, babies are often held by another person when photographed).

Simply knowing the capture conditions of an image can help identify the persons in the image. The image capture time is particularly relevant, as it allows us to group multiple images in the collection captured at the same event into clusters. Within an event, it is likely that a given person will maintain a constant appearance and wear the same clothing. The geographic location of the image capture is intuitively useful for determining the identities of people in the image.

Social context is information about people and their society that is useful for understanding images. For example, because specific first names rise and fall in popularity over time and are selected based on the gender and culture or location of the child, a first name provides prior information about the age, gender and origin of a person [33]. When multiple people appear in an image, their social relationships are related to their age, gender, and relative position within the image. The distributions of relative ages between spouses [9, 6], parents and children [20], and siblings [7] are either documented in or can be estimated from demographic statistics. A standard actuarial table [2] allow us to consider life expectancy as a prior.

Of course, each of these contextual clues are inter-related and each is known only to some degree of certainty. For example, knowing the first name of a face provides some information about the age and gender of the person. Likewise, if the age and gender are known, the uncertainty about the person's name decreases. We use probabilistic graph models to represent this uncertainty and allow all evidence to be considered.

Pixel Context	Camera Context	Social Context
Clothing	Image capture time	First name
Other people	Flash Fire	Age and Gender
Relative pose	Brightness	Social relationship
Posture	Location	Height and Weight
Glasses, hats		National origin

Table 1: Different types of context are useful for recognizing people. Items in green indicate contextual items discussed in this paper. Items in black represent ongoing or future efforts.

3 Related Work

A recent thrust in computer vision concerns the use of context in object detection and recognition. For example, Boutell and Luo use image capture metadata and timestamp to classify images as either indoor or outdoor [19]. Hoiem *et al.*[17], and Torralba and Sinha [32] describe the context (in 3D and 2D, respectively) of a scene and the relationship between context and object detection. We observe that the context in which an image is captured extends far beyond the pixel values in the image itself. Geographic location, cultural influences, and time all affect the likelihood that specific objects will appear in images. For a few simple examples, we would not expect to see images of airplanes captured prior to 1900, or images of snow in Panama. Further, Singhal *et al.*[29] demonstrate that learning the co-occurrence and relative co-locations of objects improves object recognition. Other researchers have extended the idea of considering relative location [13, 24, 27, 37] for object recognition by integrating this context into graphical models. We extend this line of work by exploring the contribution of various types of context for recognizing people.

The most popular method for recognizing images of people is face recognition. There are many techniques for recognizing faces, or for comparing the similarity of two faces [40], and under controlled environments, recognition rates exceed 90% [25]. However, there are significant differences between the problem of face recognition in general and the problem we are addressing. As illustrated in Figure 1, a face of unknown identity is compared against a gallery of face images with known identity, where each gallery image is captured with similar pose, illumination and expression [16, 23]. For individual consumers, developing such a gallery of the subjects in their images is inconvenient at best and impossible at worst. Researchers have incorporated face recognition techniques to aid searching, retrieving, and labeling of consumer images [1, 15, 35, 39]. All of these systems rely on the user to label example faces for each individual to be recognized.

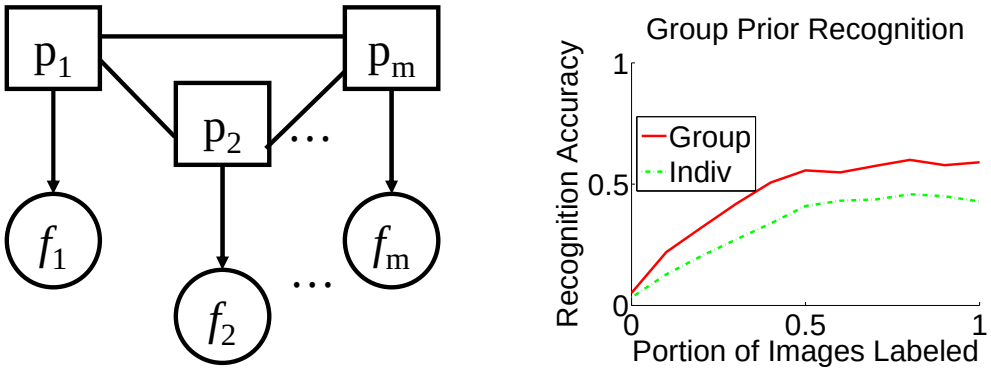


Figure 3: **Left:** A graphical model that represents the appearance features f and the identities people p in an image. Each person p_m has an undirected connection to all other people. **Right:** The accuracy of recognition improves with the group prior. The accuracy of recognition is shown as a function of the portion of the test image collection that is labeled.

Several researchers have attempted to recognize people from contextual information that extends beyond pixel data. In an extreme example, Naaman *et al.*[21] describe an interactive labeling application that uses only context (e.g. popularity, co-occurrence, and geographic re-occurrence) to create a short drop-down list for labeling the identities of people in the image. This method uses no image features, although the authors note that the combination of context- and content-based techniques would be desirable.

Our contributions are to use a data-driven approach to understanding images of people. Context from images, cameras, and demographic statistics are merged into probabilistic models for understanding images of people. This contextual data is a substitute for the intuition that a human brings to the task of image understanding.

4 Context Aided Recognition Examples

In this section, several case examples are presented which demonstrate the benefits and methods for combining context with appearance for understanding and recognizing images of people.

4.1 The Group Prior

In [12], a group prior is used to learn social groups that well-explain the observed image facial features of groups of people in consumer image collections. Rather than inferring the identity of each person in an image in a vacuum, the identities of all people are inferred simultaneously. A good assignment is one where that social group is likely to appear together, and the appearance of each person is well-explained. Face recognition is shown to be substantially improved using the group prior across several image collections. Figure 3 graphically models the relationship between the identities of the people in the image and the observed features. The set of M people in the image is denoted $\mathbf{p}$, the set of all features is $\mathbf{f}$, and $\mathbf{n}$ is a subset of $\mathbf{N}$ with M elements and is a particular assignment of a name to each person in $\mathbf{p}$. A particular person in the image is p_m, the associated features are f_m, and the name assigned to person p_m is n_m.

The appearance features f_m are derived solely from the pixels of the face region in the image. A face detector is used to locate the position and scale of the face. Feature dimensionality is reduced by first extracting facial features using an Active Shape Model [8]. The ASM locates 82 key points including the eyes, eyebrows, nose, mouth, and face border. Following the method of [12], PCA is used to reduce the dimensionality to five features.

The joint probability $P(\mathbf{p} = \mathbf{n}|\mathbf{f})$ of all the M people in a particular image, given the set of features is written:

$$P(\mathbf{p} = \mathbf{n}|\mathbf{f}) = \frac{P(\mathbf{f}|\mathbf{p} = \mathbf{n})P(\mathbf{p} = \mathbf{n})}{P(\mathbf{f})} \tag{1}$$

$$\propto P(\mathbf{p} = \mathbf{n}) \prod_m P(f_m|p_m = n_m) \tag{2}$$

Consistent with the model, we proceed from (1) to (2) by recognizing that the appearance of a particular person f_m is independent of all other individuals in the image once the identity of the individual p_m is known to be n_m. Because the size of the group prior $P(\mathbf{p} = \mathbf{n})$ grows exponentially with the number of elements in $\mathbf{n}$ (the number of faces in an image with people of unknown identity), it is estimated from the pairwise co-occurrences of pairs of people in images as follows: Let $P(n^u, n^v)$ be the probability that persons n^u and n^v appear together in an image. This probability is estimated with MLE counts from the labeled subset of

images, with a small regularization term to enforce the possibility that any two people could in theory appear together in an image. Then,

$$P(\mathbf{n}) = \frac{\prod_{u,v \in \mathbf{n}} P(n^u, n^v)}{\sum_{\mathbf{q} \subseteq \mathbf{N}} \prod_{u,v \in \mathbf{q}} P(n^u, n^v)} \tag{3}$$

where $\mathbf{q}$ has M elements. Equation (3) represents the group prior for any number of particular people appearing together in an image as a fully connected pairwise Markov model, again consistent with our model.

To test the idea, four consumer image collections including 1084 images containing 1924 labeled instances of 114 individuals were used. Within a collection, a random subset of images were selected to be labeled. Note that at least one image is always held out for testing the recognition accuracy. The labels are ambiguous; a label indicates that a particular person is in the image but does not indicate which face belongs to which label. A modified *k*-means clustering, similar to [5, 34, 38], is performed to assign ambiguous labels to faces. These label assignments are used to learn the appearance model $P(f_m|p_m = n_m)$, represented as a multidimensional Gaussian, for each individual n_m and also learn the group prior $P(\mathbf{p} = \mathbf{n})$. Figure 3 shows the improvement in the accuracy at estimating the identity of people in unlabeled images. A correct result is one in which all of the faces in the image are correctly identified. The plot shows the result on one image collection containing 188 images with 420 instances of 5 different people. The results on other collections are similar. Estimating the group prior $P(\mathbf{p} = \mathbf{n})$, the probability that a particular group of people would appear together in an image, improves recognition in consumer image collections, as shown in Figure 3. In summary, by modeling the social relationships between the people with the group prior, we improve classification performance.

4.2 First Name, Age, Gender, and Identity

In [11], the relationship between first name, age, and gender is modeled to identify people in images. Consider Figure 4, which shows two images, each containing a pair of people. Given the first names of the people in each image, most people familiar with American first names will be able to correctly assign the first names to all four faces. If the names were merely labels that contain no information, we would expect to properly assign only two names to the correct people (by random chance). Yet humans gain an understanding of their

Figure 4: Contextual information related to first names can be used to recognize people in consumer images. (Left) An image of Sierra and Patrick. By recognizing the gender of the people and names, we can confidently conclude that Patrick must be the man on the right, while Sierra is the woman. (Right) This image contains Mildred and Lisa. Mildred, a first name popular in the early 20^{th} century, is the older woman on the right, while Lisa is the younger woman on the left. This recognition is possible for humans because of their extensive cultural training.

culture that allows them to easily perform complex recognition tasks such as illustrated here. Specifically, humans learn to associate first names with appearance, age, and gender. The apparent age or gender affects the likelihood that a person has a particular name. Likewise, a person's first name allows us to better estimate their age and gender.

In this work, the U.S. Social Security baby name database [33] is used to provide social context. This database contains the 1000 most popular male and female baby names (among applicants for a U.S. Social Security Number) for each year between 1880 and 2006 and represents over 280 million named babies. Using this data, we compute statistics related to distributions over birth year, gender, and first name as shown for example in Figure 5. First names convey a great deal of information about year of birth. Names such as "Aiden", "Caden", "Camryn", "Jaiden", "Nevaeh", "Serenity", and "Zoey" all have expected birth years more recent than 2001. Therefore, we expect recent images of people with these names to be small children. Other names experience cyclical or level popularity, and consequently reveal less about the age of the individual.

Image-based classifiers are used to estimate the age a [14, 18] and gender g [3, 36] of each person in the image. A graphical model, shown in Figure 6 is used to select the most likely assignment of first names to faces. We make the simplifying assumption that given a first name, birth year and gender are independent.

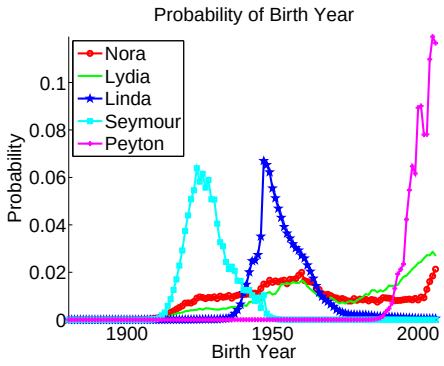


Figure 5: The distribution over birth year for a selection of first names, given the person is alive in 2007.

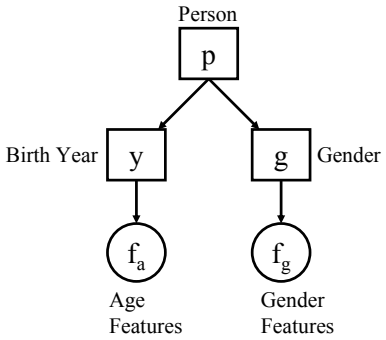


Figure 6: A graphical model that represents the relationship between a person p having a first name n, the descriptors of birth year y and gender g, and the image-based features f_a and f_g.

Appearance features related to age f_a and gender f_g are observed in the image, and we want to find the likelihood of a particular first name given these descriptor-specific features. The joint distribution can be written:

$$P(p, y, g|f_a, f_g) = P(p)P(y|p)\frac{P(y|f_a)}{P(y)}P(g|p)\frac{P(g|f_g)}{P(g)} \tag{4}$$

The term $P(g|p = n)$ is the probability that person with first name n has a particular gender. The term $P(y|p = n)$ is the probability that person with first name n was born in a particular year y. Note that we use the terms "age" and "birth year" synonymously because each conveys the same information, given that the age is known with respect to a reference year. When the image has the associated image capture time stored

in the EXIF header, the relationship between $P(a|f_g)$ and $P(y|f_g)$ is simply:

$$P(y|f_a) = P(a = c - y|f_a) \tag{5}$$

where c represents the image capture year, y represents a possible birth year and a is the age of the person. The distributions $P(g|p = n)$ and $P(y|p = n)$ are estimated from the name data, while considering the life expectancy. Finding the likelihood $P(p = n|f)$ of a particular name assignment $p = n$ given all the features $f = \{f_a, f_g\}$ is accomplished by marginalizing the joint distribution over all possible assignments of birth year and gender.

In each case, the appearance features f_a and f_g are based on the pixel values in the face region. Each face is normalized in scale (49×61 pixels) and projected onto a 37- dimensional set of Fisherfaces [4] created from an independent set of faces. For training the age classifier, image collections from three consumers with 2855 instances of 117 unique individuals were used. The birth year of each individual is provided by the collection owner, and the age of each face is found by considering birth date and image capture date. The nearest neighbors (we use 25) of a query face are found in the projection subspace. The estimated age of the query face is the median of the ages of these neighbors. Given this estimate for the age, we model $P(a|f_a)$ as a Gaussian about the estimated age, with a standard deviation of one-third the estimated age (the accuracy of our age classifier decreases with age).

Following the example of [36], we implement a face gender classifier using a support vector machine. The gender features f_g that are used are the same as the appearance features f_m from Section 4.1, that is $f_g = f_m$. A training set of 3546 gender-labeled faces from our consumer image database is used to learn a support vector machine that outputs probabilistic density estimates for gender $P(g|f_g)$.

To demonstrate the benefit provided by using first names as social context, we created a set of 148 images of 339 people by generating random pairs of first names and searching for images containing both of those named people on Flickr. The task is this: considering only a single image with detected faces and the first names of the persons in the image, determine the assignment of names to faces. Table 2 shows the results for name assignment accuracy. The complete model, utilizing both image-based age and gender classifiers, performs best. In fact, human performance on this difficult name-assignment task ranged from 67.6% to

	Overall
Random	43.7%
Age	49.0%
Gender	59.0%
Age+Gender	**61.7%**

Table 2: Using image-based age and gender classifiers for recognizing people in a single image. The percentage of correct name assignments is reported. The "Random" row value is an expectation rather than an actual experiment. The other three rows show the performance of first name assignment using the image-based age classifier, gender classifier, or both with the proposed model.

79.4%, where performance was related to the amount of time spent in the United States. Recent arrivals to the US had less time to gain familiarity with US first names, and achieved lower scores.

A secondary task includes estimating the age and gender for each face. For each person image, we manually labeled the age and gender of the person (without looking at name information associated with the image). We compare the performance of the appearance-based classifiers with that attained using the complete model. Our image-based age classifier has a mean absolute error of 10.0 years, and 28.6% of the genders are misclassified by the image-based gender classifier. Using the model, the age estimation error is reduced by 6% to 9.4 years and the gender classification errors are reduced by 32%, from 28.6% to 19.5%. The model improves the age and gender estimates over the estimates from the image-based classifiers.

4.3 Clothing as Context

Researchers have shown that clothing can be an effective tool for recognizing people. In applications related to consumer image collections [1, 31, 35, 38, 39], clothing color features have been characterized by the correlogram of the colors in a rectangular region nearby a detected face. For assisted tagging of all faces in the collection, combining face with body features provides a 3-5% improvement over using either feature independently. However, segmenting the clothing region continues to be a challenge; all of the methods above simply extract clothing features from a box located beneath the face, although Song and Leung [31] adjust the box position based on other recognized faces and attempt to exclude flesh. In our work [10] clothing regions are accurately segmented and represented by color and texture features, serving as context to improve recognition. Again, a random subset of images are selected to label the identities of the persons

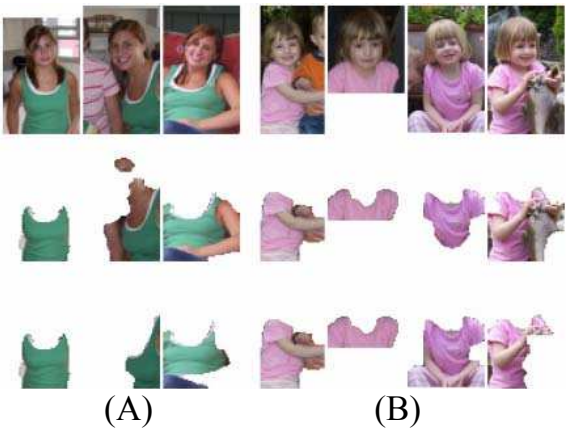


Figure 7: For each group of person images, the top row shows the resized person images, the middle row shows the result of applying graph cuts to segment clothing on each person image individually, and the bottom row shows the result of segmenting the clothing using the entire group of images. Often times, the group graph cut learns a better model for the clothing, and is able to segment out occlusions and adapt to difficult poses.

and use for training. The task is to identify the remaining people using clothing and face features. We use an example-based nearest neighbor classifier for recognizing people in this scenario.

Given an unlabeled person p, $P(p = n|\mathbf{f})$ where $\mathbf{f} = \{f, \mathbf{v}\}$ includes the facial features f and the clothing features $\mathbf{v}$, the probability that the name assigned to person p is n is estimated using nearest neighbors. Facial appearance features have been described in Section 4.1. Clothing features are extracted from the detected clothing region. The body of the person is resampled to 81$\times$49 pixels, such that the distance between the eyes (from the face detector) is 8 pixels. Texture (x- and y- derivatives) and color (luminance and two chrominance values) low-level features are extracted from the clothing region. The low-level feature vector at each pixel is quantized to the indexes of the closest visual words [30], where there is a separate visual word dictionary for color features and for texture features (each with 350 visual words). The clothing region is represented by the histogram of the color visual words and the histogram of the texture visual words within the clothing mask region. The visual word clothing features are represented as $\mathbf{v}$.

When finding the nearest neighbors (we use 9) to a query person, both the facial and clothing features are considered using the measure P_{ij}, the posterior probability that two person images p_i and p_j are the same

individual. We propose the measure of similarity P_{ij} between two person images, where:

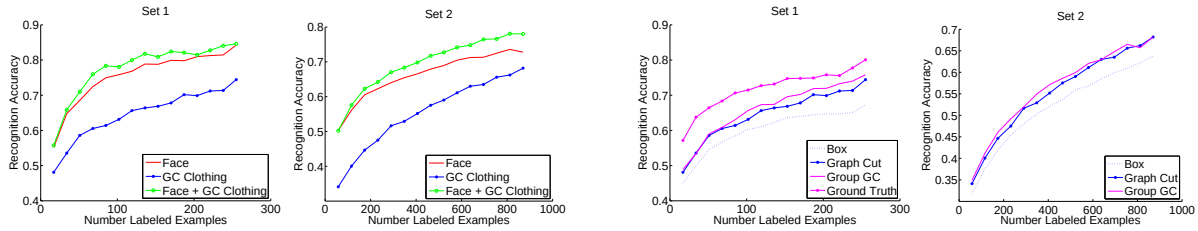
$$P_{ij} = P(S_{ij}|\mathbf{f_i}, \mathbf{f_j}, t_i, t_j) \tag{6}$$

$$\approx \max\left[P_{ij}^v, P_{ij}^f\right] \tag{7}$$

The posterior probability $P_{ij}^v = P(S_{ij}|\langle \mathbf{v}_i, \mathbf{v}_j \rangle_v, |t_i - t_j|)$ that two person images p_i and p_j are the same individual is dependent both on the distance between the clothing features $\langle \mathbf{v}_i, \mathbf{v}_j \rangle_v$ using the visual word representation, and also on the time difference $|t_i - t_j|$ between the image captures. The distance between the clothing features $\langle \mathbf{v}_i, \mathbf{v}_j \rangle_v$ for two person images p_i and p_j is simply the sum of the χ^2 distances between the texture and the color visual word histograms. The probability P_{ij}^v is approximated as a function of the distance $\langle \mathbf{v}_i, \mathbf{v}_j \rangle_v$, learned from a non-test image collection for same-day and different-day pairs of person images with the same identity, and pairs with different identities. The posterior is fit with a decaying exponential, one model for person images captured on the same day ($|t_i - t_j| = 0$), and one model for person images captured on different days ($|t_i - t_j| \neq 0$). Similarly, the probability P_{ij}^f, the probability that faces i and j are the same person, is a learned function of the distance between faces.

We justify the similarity metric P_{ij} based on our observations of how humans judge the similarity between people. If we see two person images with identical clothing from the same day, we think they are likely the same person, even if the images have such different facial expression facial expressions that a judgment on the faces is difficult. Likewise, if we have high confidence that the faces are similar, we are not dissuaded by seeing that the clothing is different (the person may have put on a sweater, we reason). Nearest neighbors to a query person are found by analyzing P_{ij} and the identities of the neighbors are used to estimate $P(p = n|\mathbf{f})$.

Clothing segmentation is improved with accurate person recognition, as improved clothing models are learned from multiple instances [26]. The clothing is successfully segmented from the image by first tessellating the image with superpixels. Clothing and background models are initialized based on a template that is positioned using the detected faces as the origin. An energy function is defined, comprised of unary terms (the cost to assign a superpixel to foreground or background) and binary terms (the cost of cutting between two adjacent superpixels is a function of the similarity of their appearance). Using graph cuts, the energy minimization is solved. The resulting clothing features are shown to be useful for effectively recognizing



(a) Combining Face and Clothing Features

(b) Improving Clothing Segmentation

Figure 8: **(a)** Combining facial and clothing features results in better recognition accuracy than using either feature independently. **(b)** Improving the accuracy of clothing segmentation improves the accuracy of person recognition. Clothing features from a box are improved upon using graph cuts, which are improved upon using multiple images for the graph cuts, which are improved upon using ground truth clothing masks.

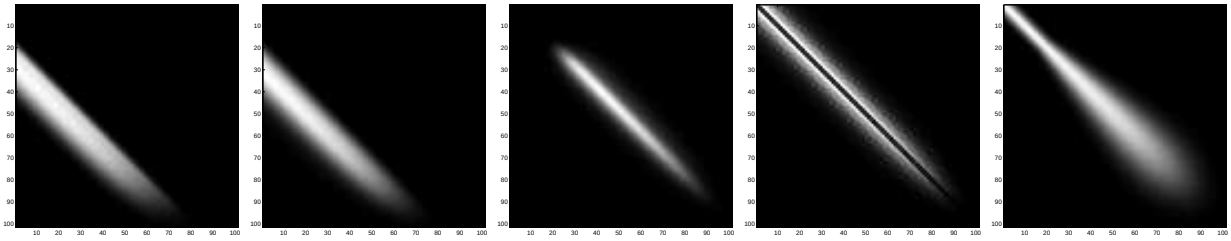


Figure 9: Each image is a representation of $P(A_1, A_2|R)$, the age of a first person and a second person sharing a social relationship R. The relationships are, from left to right: 'mother-child', 'father-child', 'wife-husband', 'siblings' and 'friends'. Except for the 'friends' relationship, all of the other joint distributions are based on demographic statistics. 'Friends' are modeled in an age-dependent fashion, as we age, we are more accepting of friends of different age.

people in consumer image collections. The clothing features along with the image capture time and facial features are merged into a single metric of similarity between two imaged persons. As shown in Figure 7, clothing segmentation is further improved with multiple examples of a person during a single event, where the person presumably wears the same clothes throughout the event.

Figure 8(a) shows that combining face and clothing features according to (6) improves people recognition accuracy for two image collections. Further, it is shown that improvements in clothing segmentation result in improvements to recognizing people (Figure 8(b)).

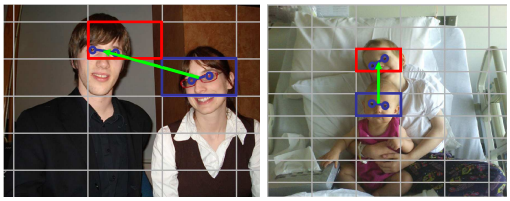


Figure 10: The relative pose between two faces is quantized into bins whose size is normalized by the average intereye distance of the pair. The quantization is finer in the vertical direction to capture the height differences between the people that provide context for our model.

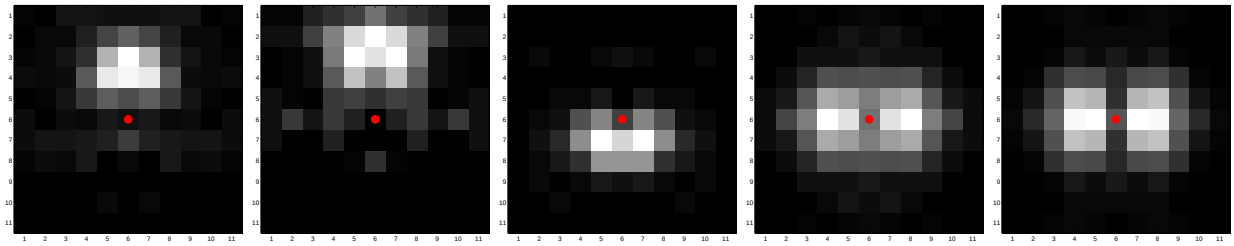


Figure 11: Each image is a representation of $P(P|R)$, the relative pose of the first person to the second person (represented by the red circle) given the pair share a social relationship R. The relationships are, from left to right: "mother-child", "father-child", "wife-husband", "siblings" and "friends". In each image, the width of a pixel represents the average distance between the eyes of a person in the image. Notice that a mother's face generally appear above the child's, but the distance is generally smaller than between a father and child. Spouses faces are particularly close to each other in images, with the wife's head generally below the husband's.

4.4 Social Relationships and Relative Pose as Context

People in consumer images are there for a reason. Generally, the people in an image either have a relationship with each other, or with the photographer. For example, if we are told that an image of a pair of women contains a mother and her daughter, we would usually be able to pick out which person is the mother and which is the daughter by ascertaining the relative ages between the pair. In fact, knowing this social relationship exists allows us to improve our age estimates for each person (since we have an intuition about the relative age differences between mother and their children). Fortunately, the characteristics of people in various social relationships are well documented by various government agencies. In this Section, we merge highly disparate context from images and from demographic databases to perform difficult inference about the relationships of people in the image.

4.4.1 Social Relationships from Demographic Statistics

Using available data, it is possible to model the distributions between the ages of people involved in different social relationships R. We consider $R \in \{\texttt{mother-child}, \texttt{father-child}, \texttt{husband-wife}, \texttt{siblings}, \texttt{friends}\}$, as shown in Figure 9, using demographic statistics from [2, 6, 7, 9, 20, 33]. The relationship between gender of a pair of people and the social relationship $P(g_i, g_j|R)$ is defined by the relationship deterministically, such that "mother" and "wife" imply $g = \texttt{female}$, "father" and "husband" imply $g = \texttt{male}$, and other relationship roles are equally likely to be male or female.

We define the relative pose between two faces as the position of the second face relative to the first. To find the relative pose k_{ij} between persons p_i and p_j in the image, the eyes of each of the pair of faces are located with an Active Shape Model [8]. The average inter-eye distance of the pair of people is found and used to normalize the coordinate system. The position of the second face relative to the first is found in this quantized normalized coordinate system. We use a rectangular quantization of 11×11 or 121 total bins, with finer quantization in the vertical dimension. Horizontally, this represents a maximum face separation of 20 inter-eye distances between the faces of the pair. In practice, the model has been found to be robust to different quantization schemes given our training data. Figure 10 illustrates the process of quantizing the relative pose of a second face (in red) with respect to a first face (in blue) and the coarseness of the quantization. Furthermore, the relative position that people take within an image is related to their relationship. For example, young babies often appear on the lap of a parent in a photograph. In a portrait of a husband and wife, the husband's head is generally beside but above the wife's due to the physical differences between males and females [22]. Learning from thousands of photographs where the relationships are known, the distribution of the relative pose of the two people can be learned for each relationship, and is shown in Figure 11.

By modeling the relationship between age, gender, pose and relationship with a graphical model (a Bayes network in this case, shown in Figure 12), it is then possible to perform inference over any of those variables using standard variable elimination. Figure 13 shows, from a test set of 148 images, the images with the highest probability of containing two specific relationships.

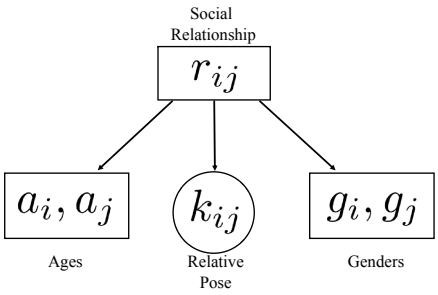


Figure 12: Social relationship r is used as a latent variable for using publicly available demographic data for learning factors that capture the interactions between relative pose, age, and gender of pairs of people.

Figure 13: **Top:** Out of the test set of 148 images, the 3 ranked as having the highest probability for containing a mother and child using the context model. The third image is an error resulting from an incorrect gender classification for the adult. **Bottom:** The 3 images ranked as having the highest probability for containing a husband and wife using the context model.

4.4.2 Relative Pose to Recognize Specific Individuals

The relative pose of specific individuals from an image collection can also be modeled as a characteristic useful for recognizing that individual. A model of the relative pose of a given individual to others in an image collection can also be used as a feature for identifying an unknown person. When identifying a query face incorporating relative pose as context, we must estimate $P(p_m = n_m | f_m, \mathbf{k_m})$, the posterior that person p_m has identity n_m, given the facial appearance f_m and the collection of relative poses of the person to others in the image $\mathbf{k_m} = \begin{bmatrix} k_{m1} & k_{m2} & \dots & k_{mM} \end{bmatrix}^{\mathrm{T}}$. We assume independence between facial appearance and

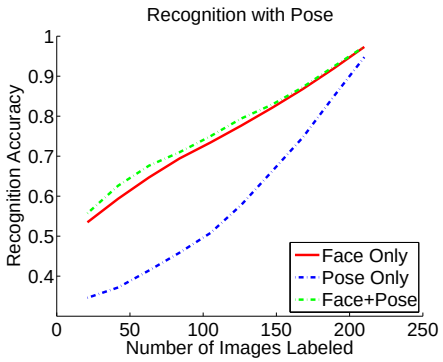


Figure 14: The relative pose between people in an image is a contextual feature that improves recognition in image collections with images of multiple people. Faces are recognized from their relative pose, from facial features, or the combination of both. This image collection contains 30 unique people in 220 images, and each curve represents over 220,000 recognition events. Combining relative pose with facial features provides a small (1.5% on average) but statistically significant benefit over using facial features alone.

each pairwise relative pose element of $\mathbf{k_m}$ to approximate $P(p_m = n_m|f_m, \mathbf{k_m})$:

$$P(p_m = n_m|f_m, \mathbf{k_m}) \propto P(p_m = n_m|f_m) \prod_{i \neq m}^{M} P(k_{mi}|p_m = n_m) \tag{8}$$

Figure 14 shows the results of using relative pose to recognize specific individuals using images with multiple people from two image collections. The faces in a random subset of images are labeled and the rest are used to test the recognition accuracy. As before, $P(p_m = n_m|f_m)$ is estimated with nearest neighbors using facial appearance features as described in Section 4.1. The individual relative pose terms $P(k_{mi}|p_m = n_m)$ are learned from the labeled subset of images. To generate each data point, the test is repeated for 50 randomly labeled subsets. The results show that relative pose is a useful feature for recognizing people. The benefit from combining relative pose with facial features is greatest when the number of labeled images is small. This is perhaps expected; when appearance is not well characterized, there is more benefit to introducing additional context, corroborating the work of [24].

5 Conclusion

Images of people must be interpreted in the context of the culture in which they are captured. A good understanding of the context in which a person appears in an image provides a strong prior for image understanding. This work described our efforts to consider context to aid in the recognition of people. Context includes information regarding social groups, first names, clothing, and the juxtaposition of faces within an image. Each of these aspects of context is useful for recognizing people in consumer images. This work appears to be among the first to incorporate demographic statistical data as context for interpreting images of people.

Significant results have already been achieved by considering context such as image capture time in combination with clothing, first names, age and gender, and social groups. Remaining work includes expanding the models to consider even more elements of context, and combining the various aspects of context into a single model.

References

[1] D. Anguelov, K.-C. Lee, S. Burak, Gokturk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *Proc. CVPR*, 2007.

[2] E. Arias. United States life tables, 2003. Technical report, National Center for Health Statistics, 2006.

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

[4] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *PAMI*, 1997.

[5] 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.

[6] M. Bhrolchain. The age difference at marriage in England and Wales: a century of patterns and trends. *Population Trends*, 2005.

[7] A. Chandra, G. Martinez, W. Mosher, J. Abma, and J. Jones. Fertility, family planning, and reproductive health of U.S. women: Data from the 2002 national survey of family growth. National Center for Health Statistics, 2005.

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

[9] M. Duffy, K. Hempstead, E. Bresnitz, F. Jacobs, and J. Corzine. New Jersey health statistics, 2004. http://www.state.nj.us/health/chs/hlthstat.htm, 2007.

[10] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *Proc. CVPR*, 2008.

[11] A. Gallagher and T. Chen. Estimating age, gender, and identity using first name priors. In *Proc. CVPR*, 2008.

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

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

[14] 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.

[15] A. Girgensohn, J. Adcock, and L. Wilcox. Leveraging face recognition technology to find and organize photos. In *Proc. MIR*, 2004.

[16] R. Gross, I. Matthews, and S. Baker. Eigen light-fields and face recognition across pose. In *FGR*, 2002.

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

[18] A. Lanitis, C. Taylor, and T. Cootes. Toward automatic simulation of aging effects on face images. *PAMI*, 2002.

[19] J. Luo, M. Boutell, and C. Brown. Pictures are not taken in a vacuum - an overview of exploiting context for semantic scene content understanding. *IEEE Signal Processing Magazine*, 2006.

[20] J. Martin, B. Hamilton, P. Sutton, S. Ventura, F. Menacker, S. Kimeyer, and M. Munson. Births: Final data for 2005. Center for Disease Control, National Vital Statistics Reports, 2007.

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

[22] National Center for Health Statistics. CDC growth charts, United States. http://www.cdc.gov/nchs/data/nhanes/growthcharts/zscore/statage.xls, 2007.

[23] K. Nishino, P. Belhumeur, and S. Nayar. Using eye reflections for face recognition under varying illumination. In *Proc. ICCV*, 2005.

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

[25] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, K. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale results, 2007.

[26] C. Rother, T. Minka, A. Blake, and V. Kolomogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *Proc. CVPR*, 2004.

[27] 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.

[28] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In *Proc. ICAFGR*, May 2002.

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

[30] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003.

[31] Y. Song and T. Leung. Context-aided human recognition- clustering. In *Proc. ECCV*, 2006.

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

[33] U.S. Social Security Administration. Baby name database. http://www.socialsecurity.gov/OACT/babynames.

[34] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In *Proc. ICML*, 2001.

[35] R. X. Y. Tian, W. Liu, F. Wen, and X. Tang. A face annotation framework with partial clustering and interactive labeling. In *Proc. CVPR*, 2007.

[36] M.-H. Yang and B. Moghaddam. Support vector machines for visual gender classification. *Proc. ICPR*, 2000.

[37] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *Proc. CVPR*, 2008.

[38] L. Zhang, L. Chen, M. Li, and H. Zhang. Automated annotation of human faces in family albums. In *Proc. MM*, 2003.

[39] L. Zhang, Y. Hu, M. Li, and H. Zhang. Efficient propagation for face annotation in family albums. In *Proc. MM*, 2004.

[40] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 2003.