A Framework for Using Context to Understand Images of People

Andrew C. Gallagher

Submitted in partial fulfillment of the
requirements for the
degree of Doctor of Philosophy

Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213

May 2009

Thesis Committee:
Tsuhan Chen, Chair
Alexei A. Efros
Martial Hebert
Jiebo Luo
Marios Savvides

Abstract

When we see other humans, we can quickly make judgements regarding many aspects, including their demographic description and identity if they are familiar to us. We can answer questions related to the activities of, emotional states of, and relationships between people in an image. We draw conclusions based not just on what we see, but also from a lifetime of experience of living and interacting with other people. In this dissertation, we propose contextual features and models for understanding images of people with the objective of providing computers with access to the same contextual information that humans use.

We show through a series of visual experiments that humans can exploit contextual knowledge to understand images of people. Recognizing other people becomes easier when the full body is shown instead of just the face. Social context is exploited to assign faces to corresponding first names, and age and gender recognition is improved when subjects see a face from an image in context with the other faces from the image instead of only a single face.

In this dissertation, we provide contextual features and probabilistic frameworks to allow the computer to interpret images of people with contextual information. We propose features related to clothing, groups of associated people, relative positions of people, first name popularity, anthropometric measurements, and social relationships. The contextual features are learned from image data and from publicly available data from research organizations. By considering the context, we show improvement in a number of understanding tasks related to images of people. When applied to collections of multiple people, we show that context improves the identification of others in the collection. When considering single images, we show that context allows us to improve estimates of demographic descriptions of age and gender, as well as allowing us to determine the most likely owner of a first name such as "Taylor". Finally, we show that context allows us to perform high-level tasks such as segmenting rows of people and identifying the horizon in a single image of a group of people.

This work shows that people act in predictable ways, for example that human patterns of association contain regular structure that can be effectively modeled and learned. From a broad perspective, this work shows that by exploiting information that is learned about people (in any field of science) we can improve our understanding of images of people.

Acknowledgements

After nine years working for Eastman Kodak, I decided to pursue a Ph.D. at Carnegie Mellon. This decision was at the same time the most exciting, and the most frightening of my career. It also turns out to have been the best decision I could have made. Carnegie Mellon is really a special place and I thoroughly enjoyed my time there. I am so thankful for the opportunity I had to study with such a creative group of students and faculty.

I am thankful for my advisor, Tsuhan Chen. As soon as our first conversation, he presented a clear path to overcoming obstacles to achieve my goal of studying at CMU. In the years since, I have seen this first impression of Prof. Chen is accurate; he is one who finds the way to accomplish a goal rather than dwelling on a problem. I've enjoyed my interactions with many CMU faculty members, including those on my committee. They collectively possess a combination of enthusiasm, intuition, and rigor that make research so much fun and so challenging. I can trace specific decisions I've made in this work to the influence of each of the committee members.

I thank my Kodak collegues and supervisors for accomodating my plan to pursue academic studies. In particular, David Cok went beyond the call of duty, and I am grateful to him.

I also thank the people who allowed me to use their image collections for my research, including Michael Ockrin, Avik Ganguly, Matthew Benton, and Alex Loui and everyone who posts images to Flickr.

I thank my parents Michael and Marie Gallagher for laying the groundwork years ago by encouraging inventive ideas as a child, whether they involved melting pennies, making random contraptions, or building humane mousetraps. I thank my super wife Holly for her support and for her example as we embarked on this great adventure with three small children. Neither of us knew quite what we were getting into when I filled out the admission form in 2005. I'm so impressed with how she used this time to learn and teach so many new things and stayed flexible the whole time. She'll easily win the "Wife of the Year" yet again for 2009! My children Hannah, Jonah, and Ethan are my inspirations, and I look forward to many years of "inventions" with them!

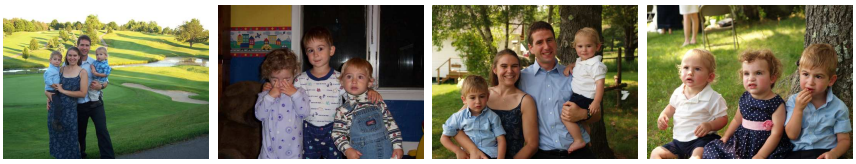
Contents

Chapter 1

Introduction

Images of people are of particular importance to the field of computer vision, relevant to both the domains of consumer imaging and security. With only a glance at an image, even small children can recognize people as well as comprehend the story behind the image. People accomplish these amazing understanding feats in part because they have the ability to interpret the image based on context. While computer vision is presently not nearly as capable, this ability defines the overarching goal to which this thesis contributes: to understand images of people with context. We want to describe or recognize people and their activities and associations from images.

An image contains a great deal of information related to physical entities such as objects and surfaces. However, we must also recognize that because of the role that photography plays in human society, an image also contain information related very much to the behaviors of humans themselves. Consider the image in Figure 1.1(a). We can easily see that is an image of four people, one woman, one man and and two children. Beyond the simple description of the people in the scene, it is also reasonable for us to surmise that this is an image of a family comprised of a mother, a father, and two children. How were we able to jump to this conclusion? We carry with us a great deal of "intuition" based on our personal experiences and observations that enables us to interpret the image. We know that the people in an image or collection of images are not selected at random from the world population; rather they generally have strong social or familial bonds. Further, we know that parents are typically a few decades older than their children. Taking this contextual evidence as a whole, it is a plausible explanation that the image is of a couple and their children. This thesis presents features and a framework for considering



(a) Family (b) Children (c) Children (d) Children

FIGURE 1.1: Images with people can be extremely complex. By considering context, we reason that (a) is a family, the child on the left in (b) is a young girl (even though her face is covered), in (c) we conclude this family is the same as the one in (a), and in roughly the same positions even though the image was captured five months later. In (d), we recognize the two boys from (a), (b) and (c), and suppose that the girl might also be the same one who was occluded in (b) (she is). By considering all available contextual clues, our understanding of these images can be improved.

context in the interpretation of the images of people. The context can either be learned from images or from other statistical sources such as national health databases.

The goal of this thesis 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. In this thesis, we are interested in using context to understand images of people in two distinct but related problem domains. The first is recognizing people in collections of consumer images. In this scenario, a person captures images of family and friends. Typically, collections contain between 20-80 distinct individuals, although certain people appear more often than others (e.g. family members). Secondly, we use context to interpret single images of people. In this scenario, we model the context associated with the image to answer questions such as: What are the ages and genders of people in the image? Which face in this image is most likely to be "Taylor"? How many rows of people are in the image? And even, is this a picture of people dining?

1.1 What is Context?

Context is broadly defined as information relevant to something under consideration. In [134], the definition of context in computer vision is *information relevant to the detection task but not directly due to the physical appearance of the object*. In our work, context includes 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 on understanding images of people. For a good summary of the types of context that have been considered in the field of computer vision as a whole, see [36].

In this work we explore the use of several classes of context in applications related to understanding images of people. The classes we define are *pixel context*, *capture context*, and *social context*. Pixel context includes context derived from analyis of non-face image regions. For example, distinctive clothing 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 [128]. 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 [14, 39], parents and children [83], and siblings [24] are either documented in or can be estimated from demographic statistics. A standard actuarial table [5] allow us to consider life expectancy as a prior.

Of course, each of these contextual clues are inter-related and each may be 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.

1.2 Contributions

The key contributions of this thesis are as follows:

Pixel Context	Capture Context	Social Context
Clothing	Image capture time	First name
Other people	Location	Age and Gender
Relative pose	Calibration Parameters	Social relationship
Posture	Flash Fire	Anthropometric Data
Glasses, hats		Personal Calendar

TABLE 1.1: Different types of context are useful for recognizing people. Items in green indicate contextual items discussed in this paper. Items in black represent potential contextual features for future work.

- The introduction of novel contextual features for improving person recognition and demographic descriptions from images of people, including novel clothing features and segmentation, the group prior for describing associations between people, and the relative pose between people.

- We show that understanding images of people improves by considering social context, a context that relates to social environment in which an image is captured. Our results show that well-studied phenomena from the social sciences can be used to improve image understanding. We believe this work represents the first demonstration of using raw demographic statistics as social context to significantly improve a computer vision task. [50, 53]

- Probabilistic models for inference on images of people with factors that represent relationships between variables which can be learned from images or from demographic and anthropometric data. [50, 52]

- The creation of several data sets for evaluating our results and sharing with the vision community, including:

  - A personal image collection of 931 faces of 32 individuals in 589 images. [49]
  - A collection of 339 people from 148 images having first names drawn from a distribution of names given to babies. [50]
  - A collection of 5080 group images containing 28231 faces labeled for gender and age category. A subset of the images are labeled with additional information, including row designation (for 222 images). [51, 53, 54]

1.3 Thesis Overview

The dissertation is organized as follows:

Chapter 2 presents the related work.

Chapters 3 to 6 describe the use of context for understanding single images of people. Chapter 3 describes the general problem of understanding images of groups of people using content and context. Age and gender estimates are improved by considering context in addition to content alone. In addition, the global structure of the people in the image is used to determine a horizon, and to classify the activity of "group dining". The structure of the faces is also used to segment an image in rows in Chapter 4. In Chapter 5, we explore the problem of an image tagged with the first names of the people it contains. Context provided from first names, age, gender and relative pose is used to resolve the ambiguous labels. In Chapter 6, we address the problem of jointly estimating age, gender and height by using a calibrated camera and context related to anthropometric data.

Chapters 7, 8 and 9 address the use of using context to infer identity in collections of images with people. In Chapter 7, the emphasis is on the segmentation of clothing for use as context for identifying people in image collections. We show the results of retrieval based on clothing, and the improvement on recognition from combining clothing and face features. Chapter 8 describes the group prior (a prior over specific social groups) and its use to identify specific people in consumer images. Chapter 9 shows the results of merging multiple contextual cues together.

Finally, Chapter 10 presents the conclusions of the dissertation.

Chapter 2

Background

This thesis is focused on using context to understand images of people. As such, the work spans several areas of computer vision, specifically, facial and human image understanding and object recognition with context. In this chapter, we describe an overview of related work, and subsequent chapters contain additional topical discussion of the related work.

2.1 Context in Computer Vision

Our use of contextual features from people images is motivated by the use of context for object detection and recognition. Context is useful to capture the relationship between objects and other information (for example, other objects, time, location) relevant to the scene. For a few simple examples, we would not expect to see images of airplanes captured prior to 1900, images of ocean beaches in Kansas, or images of snow in Panama. Similarly, we expect boats to rest on a body of water rather than the sky.

There are many different features that represent context in images ([36]), and many ways of incorporating the context into frameworks for recognizing scenes or objects. There is general agreement that context helps detection tasks, although the benefit of context decreases as the clarity of the object in question increases [101, 134]. Essentially, this means that when an object can be confidently identified by its appearance, contextual evidence does not make that conclusion more clear.

Hoiem [66], and Torralba and Sinha [126] describe the context (in 3D and 2D, respectively) of a scene and the relationship between context and object detection. A holistic impression

of the image [125] provides contextual evidence of an object's presence, position, and size. 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-occurrence, relative co-locations, and scale models improves object recognition [55, 73, 101, 112, 114]. These contextual approaches are successful because the real world is highly structured, and objects are not randomly scattered throughout an image (or in space). Of course, each object in the image acts as context for others, so graphical models are a natural choice to combine the available information [122]. Similarly, we show that there is structure to the positions of people in a scene that can be modeled and used to aid our interpretation of the image.

2.2 Face Detection, Analysis, and Recognition

From the early days of computer vision research, images containing faces and people have been a topic of focus. In the 1960s and 1970, researchers began to investigate the computational recognition of human faces from images. In general, a query face is compared with a set of gallery faces with known identity to determine the identity of the query face. In the early papers, recognition was based on extracted features and facial measurements [17, 69]. Later research used the facial appearance from actual pixel values (rather than facial measurements) for recognition, implementing methods such as Eigenfaces [127] based on PCA, Gabor filter responses [133], Fisherfaces [11] based on linear discriminant analysis, and 3D models [16].

Other facial imaging tasks have been extensively addressed in the literature. Face detection [67, 110, 137] which is essentially an exercise in sliding window classification, has been refined over several decades to the point where it executes at video rates and is a feature commonly embedded into consumer electronic products such as digital cameras.

Once a face is detected, many analysis techniques have been proposed to address specific tasks. Facial features or fiducial points are identified with deformable templates [142] or with shape or appearance models [30, 31]. Pose is estimated [88] and illumination [148] is estimated and modified. The detected face is analyzed for various attributes such as facial expression [100]. Facial hair [95, 133] and glasses [133, 135] can be detected and even removed.

Facial image analysis is used to produce a demographic description of the individual. Gender classification from a facial image is a classic computer vision problem, and a wide array of

machine learning techniques including neural networks [58], support vector machines [138], and boosting [8] have been applied. In practice, all of these methods achieve roughly similar performance [82]. Age can be estimated [56, 63, 75] or the apparent age of a facial image can be modified [76] through regression or subspace methods.

2.3 People Recognition with Context

The common thread of the work in the previous Section is that personal attributes are inferred based on the facial image itself, and each face is treated as an independent problem. However, this approach is in contrast with the system that humans use to recognize people. We know that we recognize people by integrating many contextual clues such as appearance voice, odor, gait, clothing, and body shape [131]. Despite the efforts of many researchers [115], a complete model of how humans recognize people does not yet exist. There is general agreement that as a person develops, portions of the human visual system become dedicated to face recognition [43]. This observation in itself suggests that general object recognition approaches may not achieve optimal results when applied to recognition problems associated with people. In effect, humans become experts at incorporating specialized context into their decisions regarding faces and people.

There are examples of person recognition that incorporates context, generally for multimedia applications or for applications related to organizing and retrieving images based on facial identity in consumer image collections. Image or video captions are assigned to faces by respecting the constraint that a person can only appear once per image [13, 109, 145]. In movies, scripts are used as context and matched with speakers to identify characters [41, 42].

Research devoted to solving the face recognition problem for consumer image collections has begun to employ context. A context-only solution is described by Naaman [90] *et al*. where co-occurrences between individuals in labeled images are considered for reasoning about the identities of groups of people (instead of one person at a time), though features related to appearance are not considered. Generally, at a given event where images are captured, people are wearing the same clothing. Thus, clothing (and other cues such as hair) are considered as context for recognition [3, 26, 27, 98, 117, 119, 124, 144, 145]. In an interesting application, Cao *et al*. [22] show that gender can be predicted from body shape (normally considered context) with reasonable accuracy. Geographic location derived from cellular phone or a GPS device is

considered by [35, 98]. In [121], contextual features of a social network (e.g. a list of "friends") are considered with a Conditional Random Field to reason about personal identity. When attempting to identify a person from the same day as the training data for applications such as teleconferencing and surveillance, clothing is an important cue [29, 71, 91].

Recently, several publically software applications became available that allow users to either tag or recognize faces in their collection, including Picasa 3 [59], Riya [3], iPhoto [4], and EasyAlbum [34, 85, 124]. Descriptions of these algorithms show that context including event categorization and clothing features are considered. The MediAssist package [97, 98], while not publicly available, also considers contextual features related to time, clothing, and geographic location for identifying people in collections.

2.4 Social Sciences

In the social sciences, researchers study all aspects of human function in the environment. We are interested in results from these fields for two reasons. First, in the study of the human visual system, researchers have shown strong evidence of the effect of context on object and face recognition. This provides motivation for our approach of using context to better understand images of people. Second, researchers study the social behaviors of humans and quantify these behaviors. We use this statistical knowledge as context in our models to better understand images of people. A major portion of this thesis involves incorporating the discoveries from the social sciences into computer vision algorithms to improve image understanding of images of humans.

2.4.1 Human Vision as Motivation

Certainly, learnings about the human visual system are used to justify and tune approaches for image processing and computer vision algorithms. For example, image compression algorithms such as JPEG [102] exploit knowledge of the human visual system to reduce the number of bits spent encoding data that will not be missed by the viewer. The use of Gabor wavelets is often given a biological motivation [133]. The human visual system is a proof positive that vision is possible, and without it we may have either never attempted this task, or given up long ago in our attempts to interpret visual images!

Our motivation for incorporating context into the understanding of images of people from several areas of the social sciences. In neuropsychology, the role of context in human understanding has been investigated for decades (see [9] for a review). Research has shown that humans have difficult time recognizing objects that are presented out of context. Objects displayed with an appropriate context were most effectively recognized [99], but inappropriate context results in more mistakes (e.g. a mailbox incorporated a kitchen context is mistakenly interpreted as a loaf of bread.) Biederman [15] describes relations that comprise the interactions between objects in a scene including: interposition (occlusion), support, probability, position, and size. When these contextual rules are violated, object detectability in humans decreases.

Evidence shows that person recognition by humans also benefits from context. Models of person recognition in humans contain similar functional components to those of word and object recognition [21, 118] although it is surmised that different object encodings are used. The person recognition unit can be primed by the presence of specific context [20, 132]. In [132], and experiment was performed where faces were presented in context pairs. Later, faces were again shown, and the subject was required to indicate familiarity. Performance suffers when a face was presented with a face other than as it was originally, or with no context at all. Similarly, Bruce and Valentine [20] found that recognition of a face was facilitated by a short (250 ms) pre-exposure to a related face (e.g. Jackie Kennedy precedes JFK). Thompson *et al* [123] used images of actors in natural environments to show that context plays an expecially strong role for recognizing unfamiliar faces.

Particularly strong evidence of the role of context in person recognition is uncovered by Young, Hay and Ellis [139] when they asked 22 participants to diary "mistakes" at recognizing people in daily life. In many examples, seeing a person out of context made recognition difficult, for example, seeing a clerk from the bank on the street [118].

The role of context in recognizing people is relevant especially when face recognition skills are suboptimal. The term prosopagnosia refers to the condition where an individual cannot recognize people, even close friends or relatives, from their faces. The condition is not one of general memory, but of recognition, as the affected individuals can form new short and long term memories and recognize general object categories. The cause is believed to be due to an impairment in the face processing region of the brain. In daily life, those affected by prosopagnosia can still recognize people from context [131], such as hair [37], clothing [10], and other contextual cues such as voice [38].

2.4.2 Social Sciences as Context

In anthropology and social psychology, the topic of the spacing between people during their interactions has been thoroughly studied [2, 64]. A comfortable spacing between people depends on social relationship, social situation, gender and culture. This concept, called proxemics, is considered in architectural design [68, 129] and we suggest computer vision can benefit as well by understanding the spacings between people.

Furthermore, the first names people choose for their children and the ages of people at various milestones in life are quantified [5, 14, 24, 39, 83, 128]. In addition, the medical and military establishments have extensive anthropometric data that quantify the distributions of various measurements of the human body [44, 60, 92].

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.

2.5 Position of this Thesis

The work in this dissertation builds on the work of other researchers. We continue the research dedicated to incorporating context into computer vision by specifically addressing the understanding of images of people with context. In addition to the results from computer vision, we also describe several studies on human that were performed to better understand how we incorporate context into our own understanding of people images.

A portion of the thesis addresses recognizing people in consumer image collections with context. This can be seen as an extension of the aforementioned related work by introducing new contextual features such as new clothing segmentation and features related to the relative positions of people in an image.

In another portion of the thesis, we show that by considering context, meaningful understanding is achieved on a single image of people. In both portions, features are motivated by research in the social sciences, and we train our model with both images and data that describes the actions of people in society. Because imaging is an integral part of our society, this knowledge is useful to improve our understanding of images of people.

Chapter 3

Understanding Images Groups of People with Social Context

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.

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 7.1. 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 Chapter,

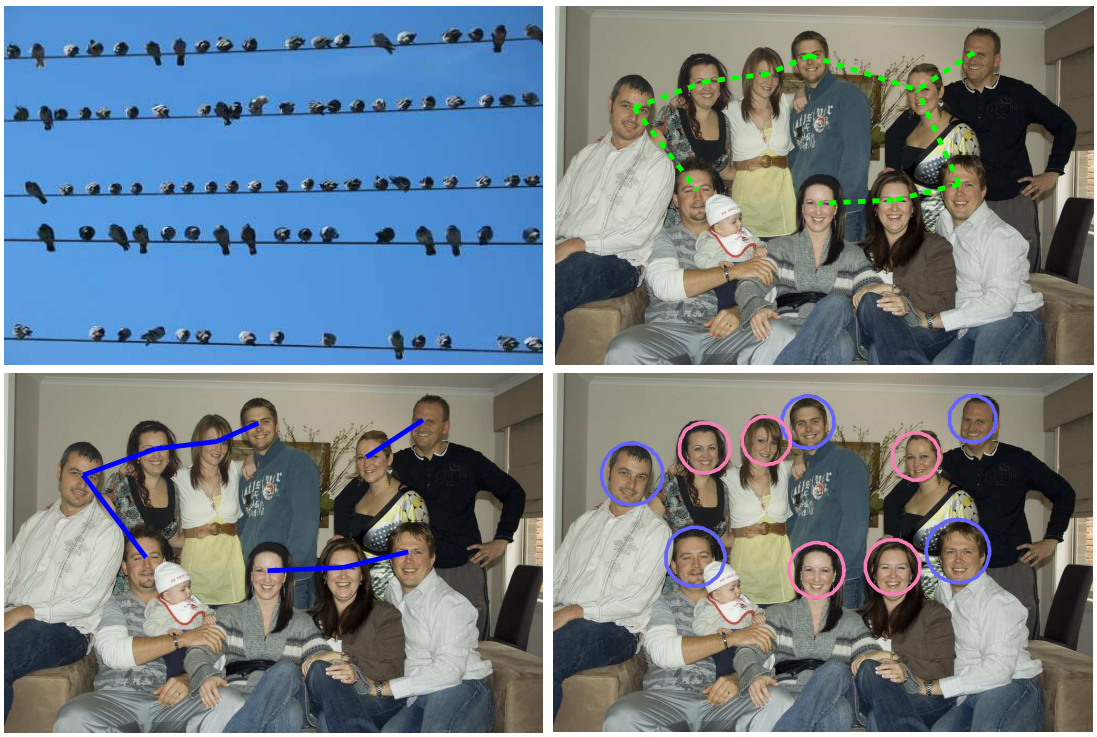


FIGURE 3.1: Just as birds naturally space themselves on a wire (Upper Left), people position themselves in a group image. We extract contextual features that capture the structure of the group of people. The nearest face (Upper Right) and minimum spanning tree (Lower Left) both capture contextual information. Among several applications, we use this context to determine the gender of the persons in the image (Lower Right).

we 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 [64] 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 [68, 129] 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.

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

3.1 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 [13], 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 [48, 90, 121], 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 [117], 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*. [66], and Torralba and Sinha [126] 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 [55, 101, 112, 114]. 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

	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 3.1: The distribution of the ages and genders of the 28231 people in our image collection.

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.

3.2 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 3.1), making this what we believe is the largest dataset of its kind [54]. 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 chapter is devoted to answering this question to the affirmative.

3.3 Contextual Features from People Images

A face detector and an Active Shape Model [30] 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 $\mathbf{r} = \begin{bmatrix} x_r & y_r \end{bmatrix}^T$:

$$\mathbf{p} = \frac{1}{2}\mathbf{l} + \frac{1}{2}\mathbf{r} \tag{3.1}$$

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.

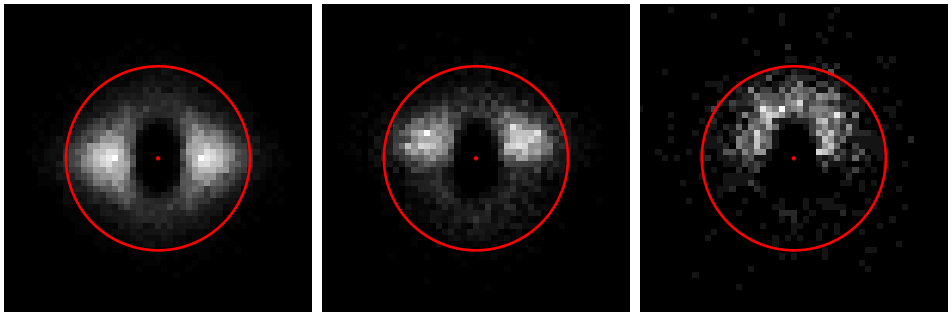
Relative Features: The centroid of all the faces in an image is found. Then, the relative position of a particular face is the position of the face to the centroid, normalized to the mean face size:

$$\mathbf{r} = \frac{\mathbf{p} - \mathbf{p}_\mu}{e_\mu} \tag{3.2}$$

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:

$$e_r = \frac{e}{e_\mu} \tag{3.3}$$

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 (3.9), the predicted size of the face compared with the actual face size is the



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

FIGURE 3.2: 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%.

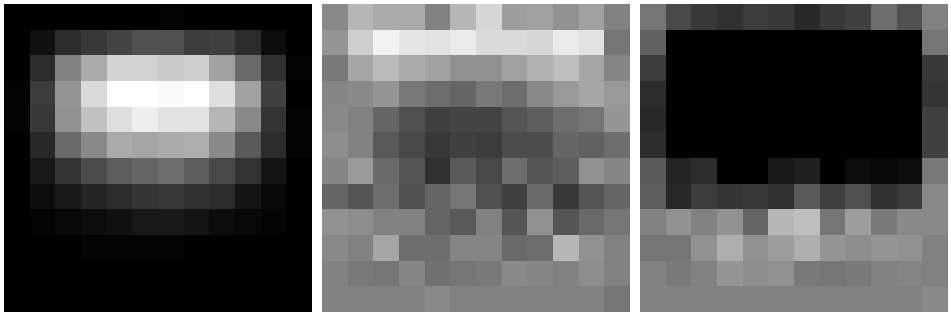
last feature:

$$e_p = \frac{e}{\alpha_1 y_i + \alpha_2} \tag{3.4}$$

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.

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.



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

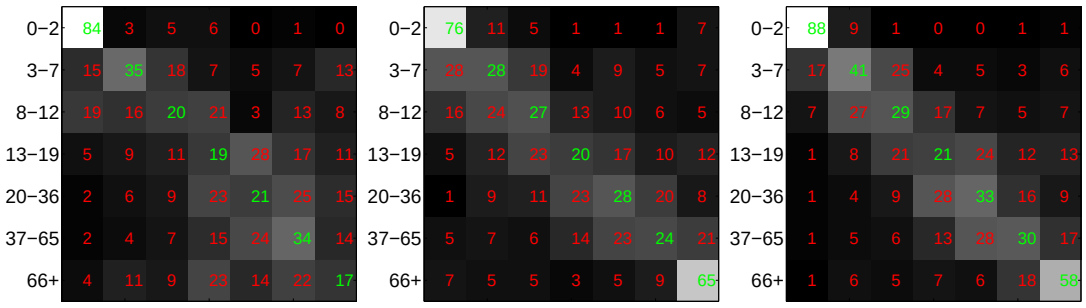
FIGURE 3.3: The absolute position of a face in the image provides clues about age and gender. Each of the three images represent a normalized image. (a) The density of all 28231 faces in the collection. (b) $P(f_g = \mathtt{male}|\mathbf{p})$. A face near the image edge or top is likely to be male. (c) $P(f_a < 8|\mathbf{p})$. A face near the bottom is likely to be a child.

(a) Random Faces (b) Average Faces

FIGURE 3.4: For each quantized (10×10) position bin in the family image subset of the group images, a random face is selected in (a). In (b), the mean of all faces is computed at each quantized position. A face near the image edge or top is likely to be male. A face near the bottom is likely to be a child.

3.3.1 Evidence of Social Context

It is evident the contextual feature $\mathbf{f}_x$ captures information related to demographics. Figure 4.3 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 ± 3 mm [44], 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



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

FIGURE 3.5: 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.

need for space for the sake of capturing an image.

Figure 3.3 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. Samples and averages of the faces at particular normailzed postions within the images are shown in 3.4. 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 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%.

3.3.2 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.

3.3.2.1 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

	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 3.2: 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).

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 3.2 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.

3.3.2.2 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 [56, 63] where a low dimension manifold for the age data is learned. 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:

$$d_{ij} = (\mathbf{f}_i - \mathbf{f}_j)\mathbf{W}\mathbf{W}^T(\mathbf{f}_i - \mathbf{f}_j)^T \tag{3.5}$$

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 is correct about 70% of the time. This is lower than others [8], 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:

$$P(f_g|\mathbf{f}_a, \mathbf{f}_x) = P(f_g|\mathbf{f}_a)P(f_g|\mathbf{f}_x) \tag{3.6}$$

Table 4.1 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 6.3 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 3.4. For example,

(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 3.6: 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.

	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 3.3: 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 3.4: For smaller faces $\leq$18 pixels between eye centers, classification suffers. However, the gain provided by combine context with content increases.

context now improves gender accuracy by 6%. This corroborates [101] in that the importance of context increases as resolution decreases.

3.4 Scene Geometry and Semantics from Faces

The position of people in an image provides clues about the geometry of the scene. As shown in [81], 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*. [33] and Hoiem *et al*. [66] 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.

3.4.1 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*. [66], we have:

$$E_i = \frac{e_i Y_c}{y_i - y_o} \tag{3.7}$$

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.

FIGURE 3.7: 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 (3.8), 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.

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 (3.7) and writing in matrix form:

$$\begin{bmatrix} E_{i1} & e_{i1} \\ E_{i2} & e_{i2} \\ \dots & \dots \\ E_{iN} & e_{iN} \end{bmatrix} \begin{bmatrix} y_o \\ Y_c \end{bmatrix} = \begin{bmatrix} y_{i1}E_{i1} \\ y_{i2}E_{i2} \\ \dots \\ y_{iN}E_{iN} \end{bmatrix} \tag{3.8}$$

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 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 3.7 shows the estimated and ground truth horizons for several images, and the accuracy is reported in Table 3.5. 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 3.5: 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.

3.4.2 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 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:

$$e_i = \alpha_1 y_i + \alpha_2 \tag{3.9}$$

and then $\frac{de_i}{dy_i} = \alpha_1$. The model from (3.8) 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 3.8 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 3.9. 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 [23] 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.

FIGURE 3.8: 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).

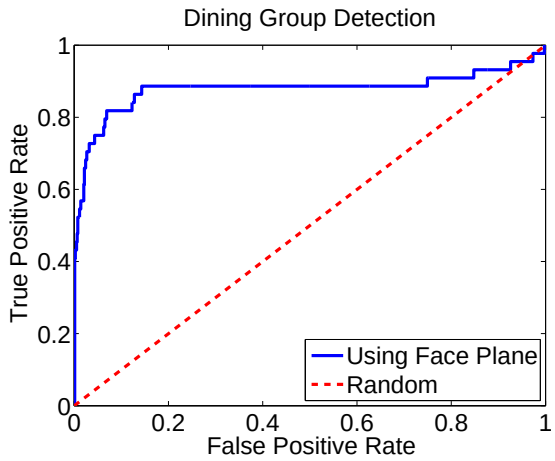


FIGURE 3.9: The face plane encapsulates semantically relevant information. The solid blue curve shows the detection of group dining images using a single feature related to the face plane. The red dashed curve shows expected random performance.

3.5 Human Understanding

In the past, human accuracy for the age recognition task has been measured [56], 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 3.10. In each stage, the subject must guess the age (in years) and gender of a face from a group

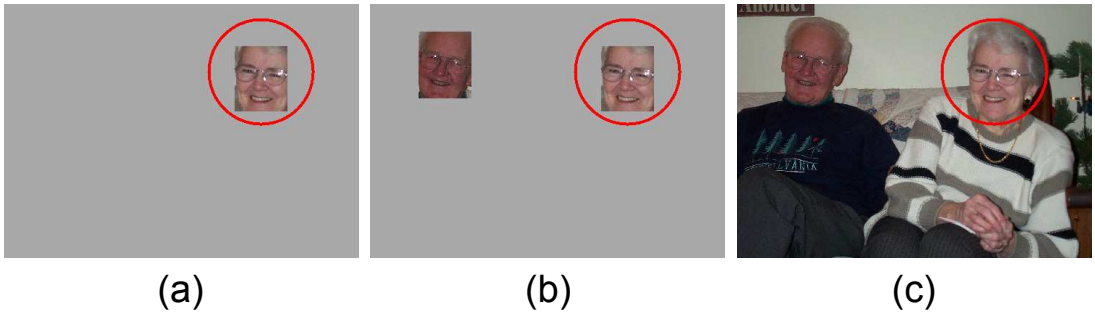


FIGURE 3.10: An example of the images shown to subjects in our human study. The subject estimates age and gender based on (a) the face alone, (b) all the faces in the image, or (c) the entire image. Human estimates of age and gender improve when additional context is available.

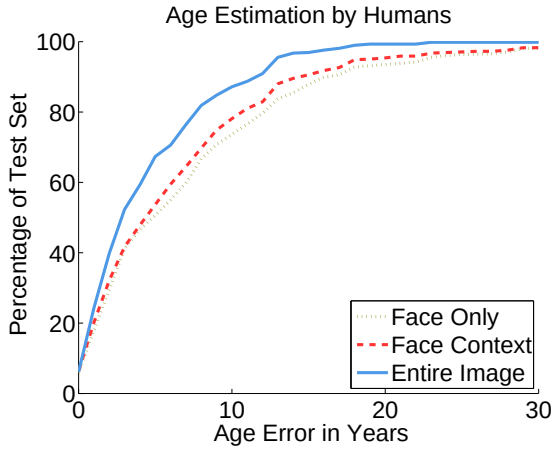


FIGURE 3.11: The effect of context on age estimation by humans. The curves show the percent of age estimates that are within a certain number of years in age error.

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.

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 3.11 and described in Table 3.6. 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

	(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 3.6: 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.

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 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.

3.6 Conclusion

In this chapter 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.

We feel this is a rich area for researchers, and we provide our image collection to other interested researchers [54].

Chapter 4

Finding Rows of People in Group Images

It is common in social gatherings to capture an image of a group of people. In these situations, people are often arranged in rows to ensure that the camera can view each face. Depending on the situation, the rows can be either highly structured, or more informal as shown in Fig. 7.1. The definition of what constitutes a row of people is not obvious. Our definition of a row of people is as follows: within a row of people, each person is at roughly constant distance from the camera, roughly in the same physical posture (e.g. sitting, standing, or kneeling), and roughly supported by the same surface (e.g. all people in a row stand on the same step in a flight of stairs). In this Chapter, we present a graph-based algorithm for detecting rows of people using graph cuts with learned energy terms.

The algorithm itself relies on the fact that there is order in the manner in which people arrange themselves in social situations. In the social sciences, the study of personal space dates to the mid-twentieth century [64]. Even without conscious effort, the relative positions of people in social situations is affected by, among other factors, age, gender, social status, the local culture, and even lighting. Our broader goal is to use the discoveries from the social sciences as *social context* for interpreting images of people. Social context is context that describes people, their culture, and the social aspects of their interactions at the time and place the image was captured.

Recovering the rows of people in a group image has applications in searching, organizing, and annotating images.

FIGURE 4.1: In many group shots, people are arranged in rows that have physical meaning in the scene. Sometimes these rows are highly structured (top) and other times less so (bottom). Our algorithm discovers rows of faces in the images. In the images on the right, each row's faces are marked with a dot of the same color.

4.1 Related Work

There is, of course, a large body of work on facial features and recognition, e.g. [103, 147]. However, the majority of this work considers each face as an independent problem. Exceptions include efforts to characterize the frequency of an individual appearing in a personal image collection and modeling the likelihood that specific combinations of people will appear together in an image or event [48, 90, 121]. None of this work considers the position of a face within the image.

Despite the prevalence of group shots, there is surprisingly little work devoted to their analysis and understanding. In [117], the authors attempt to match people in repeated shots of the same scene. Clothing, face, and hair are considered to establish correspondences. Facial arrangement was considered in [1] as a way to measure similarity between pairs of images, but was not explored as a means for understanding a single image. In [28], a rule-based system is proposed for tagging faces based on directional cues in annotations.

Regarding the method we use to solve the problem, graph cuts are used to solve many problems in computer vision [19, 106]. In pairwise models, an energy function is composed of unary and

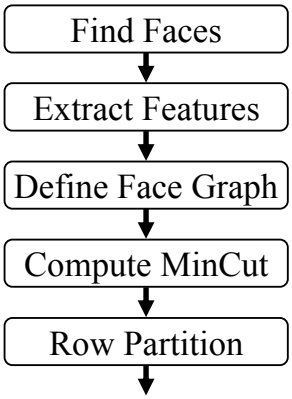


FIGURE 4.2: The flow diagram of our algorithm for finding rows of faces in an image of a group of people.

pairwise energy terms in a manner that a graph cut provides the optimal solution. However, the unary and pairwise terms are usually defined by hand (e.g., based on pixel intensity difference as in [45, 111]) to achieve good results. Recently, there have been efforts to learn distance metrics [136] from labeled training samples to better express the similarity between samples. Our problem is essentially to produce a clustering of the faces into k rows, where k is unknown. In this paper, we take the latter approach by training a classifier to distinguish between pairs of faces that are in the same row, and pairs that are in different rows. This classifier is used to establish the energy terms in our graph model.

Our contributions are the following: We present an algorithm for detecting rows of people in group images. Our approach uses graph cuts on a graph whose edge weights are learned with a classifier from training data. Our model represents the social context of personal space for solving a practical image understanding problem.

4.2 Features and the Face Graph

Fig. 4.2 shows the algorithm flow. First, faces are detected with a face detector and an Active Shape Model [30] is used to locate the left and right eye positions that serve as features. Next, an undirected graph is defined where each human face is represented as a vertex. Edge weights are learned as a function of the features of the pair of faces connected by each edge. The graph is constructed so that under certain assumptions, a minimum cut produces the most likely binary split separating the group of faces into sets of rows of faces. By recursively applying this binary split until a stopping criteria is met, the row partition of the image is found.

4.2.1 Face Features

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 $\mathbf{r} = \begin{bmatrix} x_r & y_r \end{bmatrix}^T$. The distance between the two eye center positions for the face is the size $e = \langle \mathbf{l} - \mathbf{r} \rangle_2$ 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 $\mathbf{f}_n$ as a 3-dimensional contextual feature vector $\mathbf{f}_n = \begin{bmatrix} x_n & y_n & e_n \end{bmatrix}^T$.

4.2.2 The Face Graph

Next, the face graph $G = (V, E)$ is constructed where each face 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. This graph defines a Conditional Random Field (CRF) that represents $P(\mathbf{v}|\mathbf{F})$, the probability of a labeling given the features associated with all faces in the image. We seek the most probable binary labeling $\mathbf{v}^*$ of the faces.

$$P(\mathbf{v}|\mathbf{F}) \propto \prod_n \Psi(v_n) \prod_{(v_m, v_n) \in E} \Phi(v_m, v_n) \tag{4.1}$$

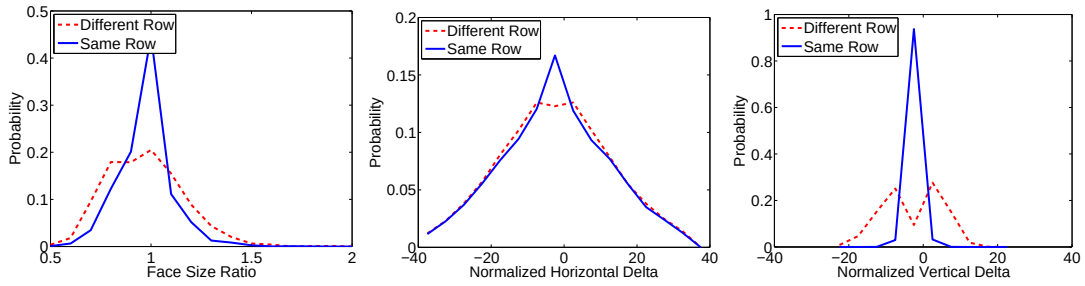
The most probable labeling $\mathbf{v}^*$ is found as:

$$\mathbf{v}^* = \underset{\mathbf{v}}{\operatorname{argmax}} P(\mathbf{v}|\mathbf{F}) \tag{4.2}$$

$$= \underset{\mathbf{v}}{\operatorname{argmin}} - \sum_n \log \Psi(v_n) - \sum_{(v_m, v_n) \in E} \log \Phi(v_m, v_n) \tag{4.3}$$

where n and m are indices over particular faces in the image. Possible row labels for each face are $v_n \in \{0, 1\}$ where 0 and 1 represent different rows. The unary term $\Psi(v_n)$ is constant because nothing in our model distinguishes between the facial features and which row that face is likely to be in. The pairwise term $-\log \Phi(v_m, v_n)$ represents the cost of assigning either the same or different labels (row indices) to a pair of faces.

From (4.3), the most likely row labeling $\mathbf{v}^*$ corresponds with the minimum cut of the graph G, when the edge weights are $-\log \Phi(v_n, v_m)$. Learning these parameters $\Phi(v_n, v_m)$ in an undirected graphical model is notoriously difficult. Intuitively we would like to be rewarded



(a) Relative Size (b) Horizontal Delta (c) Vertical Delta

FIGURE 4.3: The distributions of the features $\mathbf{f}_{mn}$ associated with a pair of faces in an image. Faces in the same row tend to be of similar size (a), and be close horizontally (b), and vertically (c).

for cutting between faces that are probably in different rows, and penalized for cutting between faces that are likely in the same row. Under the naïve Bayes assumption that each pair of faces is independent of all others, then $\Phi(v_n, v_m)$ is related to the probability $P(s_{mn}|\mathbf{f}_m, \mathbf{f}_n)$, where s_{mn} is the event that $v_m = v_n$ (faces m and n belong to the same row) as follows:

$$\Phi(v_m, v_n) = \begin{bmatrix} 1 & \frac{1-P(s_{mn})}{P(s_{mn})} \\ \frac{1-P(s_{mn})}{P(s_{mn})} & 1 \end{bmatrix} \tag{4.4}$$

Using a set of training images where each face row has been identified, the term $P(s_{mn}|\mathbf{f}_m, \mathbf{f}_n)$ can be learned with any classifier. In our work, we learn $P(s_{mn}|\mathbf{f}_m, \mathbf{f}_n)$ with Gaussian Maximum Likelihood (GML), using a multivariate Gaussian to represent each of the two classes (either $v_n = v_m$ or $v_n \neq v_m$) from the training data, where an equal prior is assumed. The feature vector $\mathbf{f}_{mn}$ is produced from $\mathbf{f}_m$ and $\mathbf{f}_n$) and represents the relative position and scales of the two faces in the image as follows:

$$\mathbf{f}_{mn} = \begin{bmatrix} \frac{e_m}{e_n} & \frac{x_m - x_n}{e_m} & \frac{y_m - y_n}{e_m} \end{bmatrix}^T \tag{4.5}$$

Fig. 4.3 shows one-dimensional projections of the distributions of the values of $\mathbf{f}_{mn}$ for each of the two class values of s_{nm}.

4.2.3 Recursive Minimum Cuts

The graph $G = (V, E)$ is formed with vertices at each face in the image and edge weights assigned according to (4.3) and (4.4). A visualization of the graph G is shown in Fig. 4.4. The

FIGURE 4.4: A visualization of the graph G for the two images from Fig. 1. Every face is a node in the graph, and edges weights are the cost to cut the edge. Green edges indicate a cost to cut (the pair is likely in the same row) and magenta edges indicate a reward for cutting (the pair is likely in different rows). Black edges neither reward nor penalize a cut.

minimum cut of the graph is found, partitioning the graph vertices into two sets V_1 and V_2. Note that the value of the cut need never exceed 0, since when V_1 is the empty set no edges are cut. The subgraph associated with each set is then recursively cut until the value of the cut is zero. In this way, the original graph G is partitioned into k components, each representing a row of people. Unlike many unsupervised clustering algorithms, the number of components does not need to be supplied by a human because the whole process is guided by the labeled training data. After the process converges, the rows are renumbered starting from the image top.

It is important to note several approximations in our approach. First, the general problem of finding a minimum cut on a graph with negative weights is NP-hard, although in special cases efficient algorithms exist. We use a spectral relaxation [111] to find an approximate solution. Let $\mathbf{A}$ represent the adjacency matrix, where each element $a_{mn} = -\log \frac{1-P(s_{mn}|\mathbf{f}_{mn})}{P(s_{mn}|\mathbf{f}_{mn})}$ is the cost associated with cutting an edge. Then the eigenvector of the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$ associated with the smallest eigenvalue is binarized to approximate the minimum cut solution. $\mathbf{D}$ is a diagonal matrix with each element equal to the corresponding row sum of $\mathbf{A}$.

Second, is must be noted that even if each recursive minimum cut is exact, there is no guarentee that the final partition will be equal to that achieved by performing an optimal exact k-way minimum cut of the graph. In image segmentation, this problem is addressed with an application of k-means after a dimensionality reduction [94]. We leave this topic as future work to explore.

Despite these approximations, the model provides useful solutions to the row segmentation problem, as shown in the next Section.

FIGURE 4.5: **Top:** Examples where our algorithm perfectly recovered the rows of people. Notice the variety of postures and arrangements of the people, from standing, sitting, and even laying. **Bottom:** Imperfect results. Sometimes people smaller (left) or taller (second image) than the rest of the row cause mistakes. Each of these images actually has only a single row of people. In the third and fourth images, mistakes occur near the right side where the algorithm is confused by a junction of multiple rows. Best viewed in color.

4.3 Experiments

To test our ideas, we collected a images from Flickr using the search string:

"`group shot`" or "`group photo`" or "`group portrait`"

The rows of people were manually labeled in 234 images. In total, these images contain 2222 faces and 465 rows of people (approximately 2 rows per image and 4.8 people per row. The number of people in each image ranges from 4 to 28, and the number of rows per image ranges from 1 to 5.

We test on one image at a time, leaving the rest of the images for training the GML classifier. We use a complete graph G over the face vertices to find the rows of people. The row clustering quality is compared to the manually labeled rows using the Rand Index [105]. Each partition is viewed as a collection of $n * (n-1)/2$ decisions, one decision per pair of data points. In a given partition of the data, two data points are either in the same or in different cluster. The Rand Index quantifies the proportion of the decisions for which the algorithm's decision and the ground truth decision match. A perfect score in this metric is 1.0, or 100%.

Table 1 4.1 reports our results. The algorithm's Rand Index is 92.6%. The algorithm achieved perfect row segmentation on greater than two-thirds of the images (67.5%). The discovered

	Accuracy
Rand Index	92.6%
Correct Images	67.5%
Correct No. Rows	73.5%

TABLE 4.1: Quantitative results from applying our algorithm to 234 images containing 2222 people.

number of rows k is correct 73.5% of the time. Fig. 4.5 provides discussion of the algorithm results on eight image examples.

4.4 Conclusions

In this paper we introduce a graph-based algorithm for finding rows of people in group images. In our approach, a graph is constructed whose minimum cut corresponds to a separation between rows of people. Our approach is shown to be effective by testing on a large number of images of people. We demonstrate that image understanding benefits by considering the social context provided by the structure of multiple people in an image. We feel this is a rich area for researchers, and we provide our image collection to other interested researchers [54].

Chapter 5

Estimating Age, Gender and Identity using First Name Priors,

In this chaper, we introduce a probabilistic graphical model that represents the relationship between age, gender, identity, first names, and relative pose and demonstrates the power of using social context for interpreting images of people. The model parameters are learned from a combination of images and large volumes of publicly available demographic data.

We combine image-based gender and age classifiers with the social context information provided by first names and relative pose to recognize people with no labeled examples. Our model uses image-based age and gender estimates along with consideration of pose for assigning first names to people and in turn, the age and gender estimates are improved. In summary, we show that social context provides valuable information for understanding images of people.

The recognition of people in consumer images is far more than solely a face recognition problem. To best understand the semantics of who is in an image, we need to understand *social context*, a context that describes people, their culture, and the social aspects of their interactions at the time and place the image was captured. For example, it has been shown that context (from clothing) assists even humans at recognize images of people [49]. To further illustrate this point, consider Figure 7.1, 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 (e.g. persons A and B), we would expect to properly assign only two names to the correct people (by random chance). Instead, we recognize that first names have semantic

FIGURE 5.1: Is it possible to recognize people for which no labeled examples exist? (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 knowledge of social context provided by first name semantics.

meaning that provide contextual information about a person, including birth year and gender. Through a lifetime of experience, humans gain an understanding of the social context provided by first name semantics that allows them to easily perform complex recognition tasks such as illustrated here.

Figure 5.2 presents a second example that illustrates the contextual information that relative pose between people provides for the interpretation of images of people. Two stereotypical images are shown. On the left is an image of a male-female couple. As might be anticipated, the physically taller male's head [92] is located higher in the image and to the side of the female's head. On the right, a small baby is sitting on the lap of an older sibling, and the heads are positioned such that the baby's is below the sibling's. In fact, just the relative positions of the heads (omitting facial detail) provides a good deal of information for our interpretation of who might be in these images. In essence, the goal is to provide the computer with the same knowledge of social context that humans use for analyzing images of people. Fortunately, rather than relying on a lifetime of experience, social context can often be modeled with large amounts of publicly available data.

With these examples in mind, it is clear that the context considered for understanding an image must extend beyond the borders of the image itself. Our model uses appearance from within the image as well as social context (the relationships between age, gender, relative pose and first name). 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. Meanwhile,

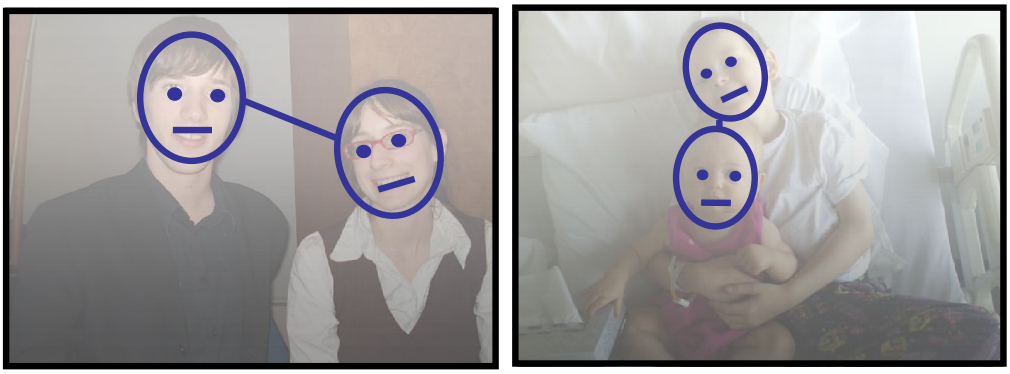


FIGURE 5.2: When multiple people are in an image, their relative pose provides clues about their age and gender. (Left) When a male and female couple are in an image, the male's head is usually above and beside the female's as a result of physical differences. (Right) A baby is often pictured on the lap of an older sibling or parent.

age and gender affect the relative positions of people in consumer snapshot images. We model and exploit these associations in our work.

This chapter makes the following contributions: First, we propose a probabilistic graphical model that captures the inter-relationships between the social context, appearance, and identity and allows for inference over these variables in a principled manner. Second, we demonstrate that the model parameters can be learned from image data as well as from large public demographic datasets that were not collected with computer vision applications in mind. Our model successfully employs this (non-image) demographic information for recognizing people. Third, this work shows the importance of integrating social context into computer vision applications. We believe we are the first to explore using the social context provided by first names or relative pose between people and we document its contribution for improving the understanding of people images.

5.1 Related Work

Computer vision research has recently become focused on the use of context in object detection and recognition. Hoiem *et al*. [66], and Torralba and Sinha [126] describe the context (in 3D and 2D, respectively) of a scene and the relationship between context and object detection. Further, Singhal *et al*. [114] demonstrate that learning the co-occurrence and relative co-locations of objects improves object recognition. Other research has extended the idea of considering relative location [55, 73, 101, 112, 143] for object recognition integrating this context into graphical

models. We extend this line of work by exploring the contribution of relative pose for recognizing people and their physical attributes.

Regarding face recognition, there are many techniques for recognizing faces or for comparing the similarity of two faces [147], and under controlled environments, recognition rates exceed 90% [103]. However, there are significant differences between the problem of face recognition in general and the problem we are addressing. Often, 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 [61, 96]. For individual consumers, developing such a gallery is inconvenient at best and impossible at worst. Researchers have incorporated face recognition techniques to aid searching, retrieving, and labeling of consumer images [3, 57, 124, 145]. All of these systems rely on the user to label example faces for each individual to be recognized and none rely on context from first name semantics or relative pose.

Both the Satoh and Kanade [109] "Name-It" system and the "Faces and Names" work [13, 62] associate names from captions with faces from images or video. The main focus in these papers is to use a large number of images to aid in the unsupervised clustering. Similarly, in Zhang *et al*. [145], a user indicates a set of images that contain a certain person. The algorithm selects one face from each image, maximizing the similarity, and concludes these faces must be the certain person. Again, names are treated merely as labels that contribute no information to the problem solution. The desire is always to assign the same label to similar faces from different images, without incorporating context. As a result, none of these papers could resolve the problem of associating multiple names and images in single image (as readily noted in [145]).

Several researchers have attempted to recognize people from contextual information that extends beyond pixel data. In an extreme example, Naaman *et al*. [90] describe an interactive labeling application that uses only context (e.g. popularity, co-occurrence, and geographic reoccurrence) 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. In [48], 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.

In this Chapter, we built on our work [50] that explores the relationship between first names, gender, and age by incorporating relative pose between people into the probabilistic model. Our complete model combines and exploits two elements of social context (first names and relative

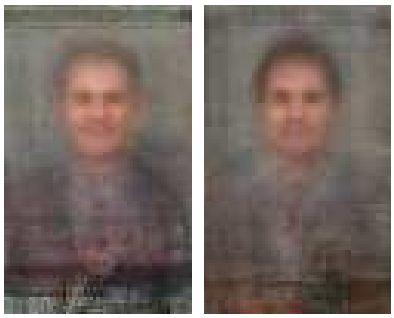


FIGURE 5.3: One approach to determining the likelihood of a first name given appearance is to learn an appearance model for each first name. For example, the figure shows the average of 50 images of people named "Tim" (Left) and "Bob" (Right). These models would need to be updated as people age and first names rise or fall in popularity. Our alternative approach is to learn the relationship between appearance and first name through the attributes of age and gender.

pose between people) that had previously been overlooked by researchers. Further, we show that learning can be accomplished using public demographic data rather than relying solely on image data.

5.2 Modeling Appearance using Social Context

Our goal is to solve the problem illustrated in Fig. 7.1, where an image is tagged with the names of the people in the image and the task is to determine the face corresponding to each name. We would like to model the relationship between the appearance f_i of a person p in an image with the first name n. Ideally, this relationship $P(f_i, p = n)$, a statistical model of appearance for each possible first name, could be learned given a huge number of training images of people and associated names. For example, by collecting hundreds or thousands of portraits of people for each possible first name, a model of the appearance of that first name could be learned. This could be an attractive approach, but it is not yet feasible for a number of reasons as described in Fig. 5.3. First, while there are billions of images of people on the internet and websites such as Flickr (www.flickr.com), it is still not easy to find images of people that have been labeled with accuracy, and a manual human review might still be necessary. Second, celebrities generally are labeled with greater accuracy but in far greater numbers than are non-celebrities. For example, a search for "Angelina" returns an inordinate number of pictures of actress Angelina Jolie, creating a sampling bias that is difficult to address. Third, this appearance model changes over time as a particular first name decreases or increases in popularity, and those already with a given first

name change in appearance as they age. Managing this evolution would be a challenging task in itself.

We take a different approach. Rather than directly learning the appearance for each name, we instead propose a set of descriptors that has an easy-to-learn relationship with both first names and the visual appearance of person images. The descriptors we select are birth year y and gender g, as we can learn $P(a|f_a)$, an image-based estimate of the person's age given age-relevant appearance features f_a and $P(g|f_g)$, the gender of the person given gender-relevant appearance features f_g. 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:

$$P(y|f_a) = P(a = c - y|f_a) \tag{5.1}$$

where c represents the image capture year, y represents a possible birth year and a is the age of the person. 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.

In our model, we consider two elements of social context, first names and relative pose between people in an image. We learn the relationship between each of these contextual items and the ages and genders of people in an image. For example, the relationship between first name, age, and birth year is contained in publicly available data. Given a first name database [128], the distributions over these same descriptors ($P(y|p = n)$ and $P(g|p = n)$), the distribution of birth years for a given first name and the distribution over gender for a given first name, are learned with maximum likelihood estimation. Relative pose represents a pair-wise term in our model, and using either a set of labeled images, or a combination of labeled images and demographic information, we learn the relationships between relative pose and the ages and genders of pairs of people in an image.

In essence, our approach amounts to the following: A first name provides a description of attributes (birth year and gender) associated with an individual. Likewise, relative pose provides information about the ages and genders of the pair of people. By extracting the image-based appearances from a person image from which these same attributes can be estimated, we can compute distributions over relative pose, name, age, and gender.

5.2.1 First Name Semantics as Context

In our work, we use the U.S. Social Security baby name database [128]. 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 (representing over 280 million named babies). The results described here could be extended to other countries and cultures given the appropriate demographic data. Using these data, we can compute statistics related to distributions over birth year, gender, and first name.

The influence of popular culture on selected names in evident in the database. For example, between 1936 and 1937, the popularity of the female name "Deanna" increased by 2000%, the largest percentage increase in the database, coinciding with the first feature length film starring popular actress Deanna Durbin in 1936. Likewise, the largest decline in name popularity occurred between 1977 and 1978, when "Farrah" fell by 78% coinciding with actress Farrah Fawcett leaving the popular show "Charlie's Angels" in 1977.

The database contains a total of 6693 unique names, with 3401 names associated with male babies, 3960 associated with female babies, and 668 shared between both genders. There is nearly twice the diversity in the names selected for females (entropy of first names, given female $H(p|g = \text{female}) = 9.20$ bits) than for males ($H(p|g = \text{male}) = 8.22$ bits). The majority of first names are strongly associated with one gender or the other. The entropy of gender is nearly one bit (0.998, less than 1.0 because male births are slightly more likely than female) but the conditional entropy of gender given first name is only $H(g|p = n) = 0.055$. However, some names are surprisingly gender-neutral. For example, the names "Peyton", "Finley", "Kris", "Kerry" and "Avery" all have nearly equal probability of being assigned to either a boy or girl.

First names also 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 stable popularity, and consequently do not reveal much about the age of the individual. For example, of all the first names, the name "Nora" leaves us with the greatest uncertainty regarding the year of birth. Figure 5.4 shows the distribution over birth year for a selection of first names, assuming that the person is alive in 2007 (when our image test set was collected). We consider life expectancy in our calculations, using a standard actuarial table [5]. We estimate there are approximately 3.9 million men named

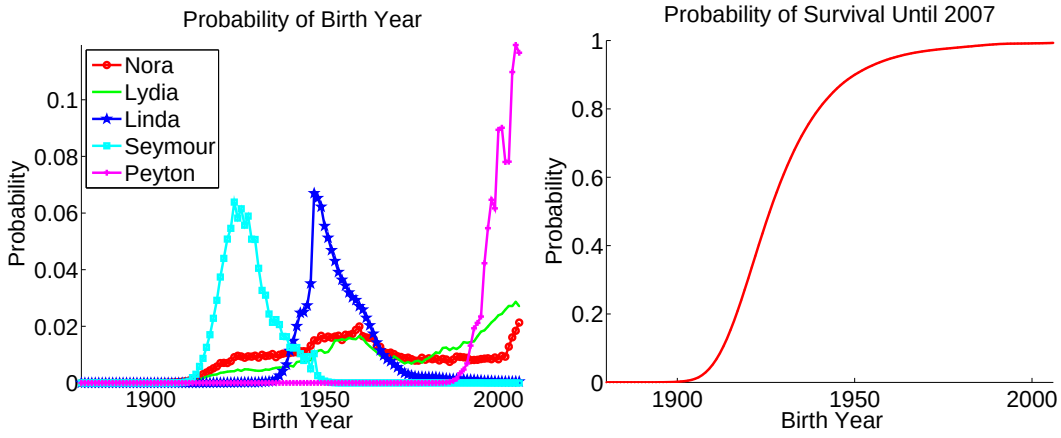


FIGURE 5.4: (Left) The distribution over birth year for a selection of first names, given the person is alive in 2007. (Right) Considering life expectancy, the probability that a person from a given birth year is alive in 2007.

"James" and 2.6 million women named "Mary" alive today in the U.S.; the most popular names for each gender.

5.2.2 Relative Pose as Context

The term "relative pose" refers to the juxtaposition of faces within in image, rather than the pose of a specific head or face in the image. The relative pose between people in an image provides a great deal of social context. Often, a vertical differential between a pair of faces in an image provides insight into the relative heights of each person, which in turn provides information about gender and age. Further, the horizontal displacement between a pair of faces in an image indicates how physically close the two are, and often tells us something about the social relationship that the two share. Because the vertical and horizontal dimensions each have a relevant semantic interpretation, we maintain the rectangular coordinate system (rather than polar) when quantizing the relative pose of a pair of people.

People in consumer images are there for a reason. Generally, the people in an image share some kind of social relationship with each other. 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 know something about the relative age differences between mother and

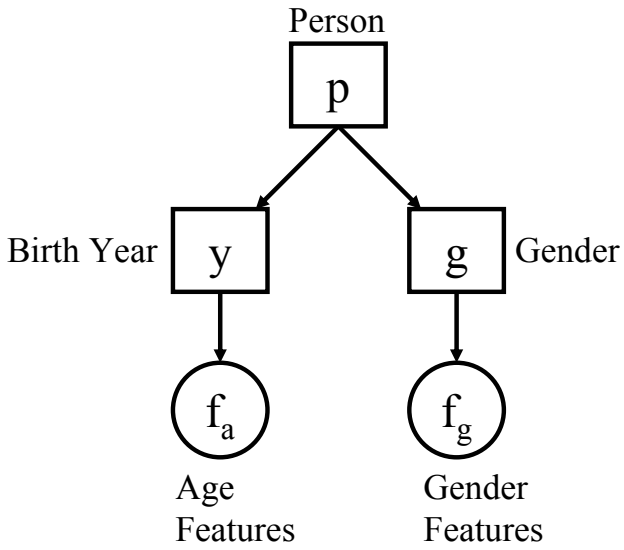


FIGURE 5.5: A graphical model that represents the relationship between a person p having a first name n, the descriptors of birth year y and gender g, and the image-based features f_a and f_g.

their children). Fortunately, the characteristics of people in various social relationships are well documented by various government agencies.

Using available data, it is possible to model the distributions between the ages of people involved in different social relationships, as shown in Figure 5.11, using demographic statistics from sources as described in Section 5.4. In our work, we have two ways to learn about pose: Purely from images, or from a combination of images and demographic data.

5.3 Social Context Probabilistic Models

In this section, we introduce probabilistic graphical models to represent the relationships between people in images and the social contexts of first names and relative pose. Of course, each of these contextual clues are inter-related and each is known only to some degree of certainty. For example, knowing the 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 models to represent this uncertainty and allow all evidence to be considered.

The graph models allow us to infer the names of people in an image, based on the beliefs regarding the ages and genders of the people. Each model asserts independence between the names

of the people in the image and their appearance, given the attributes of ages and genders. Intuitively, this assumption means that once the age and the gender of a person are known, the appearance features provide no new information about the identity of the person. For convenience, we quantize all variables, age with 101 bins each representing a year, gender with 2 bins, and relative pose with 121 bins.

5.3.1 One Person

For a person in an image, we extract image-based features related to each of the descriptors (gender and age). The name of the person and the values of the descriptors are represented as random variables. We make the simplifying assumption that given a first name, birth year and gender are independent, as in the graph model of Figure 5.5. Appearance features related to birth year and 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 is written:

$$P(p, y, g|f_a, f_g) = P(p)P(y|p)\frac{P(y|f_a)}{P(y)}P(g|p)\frac{P(g|f_g)}{P(g)} \tag{5.2}$$

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 p with first name n was born in a particular year. This distribution is estimated from the name data, while considering the life expectancy as follows:

$$P(y = i|p = n, c) \propto \text{count}(y = i, p = n)_{c-i}p_0 \tag{5.3}$$

where the notation ${}_{c-i}p_0$ is used in actuarial science to indicate the probability of survival from birth (age 0) to age $c - i$, where c is the image capture year (since we know the person is alive in this year).

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.

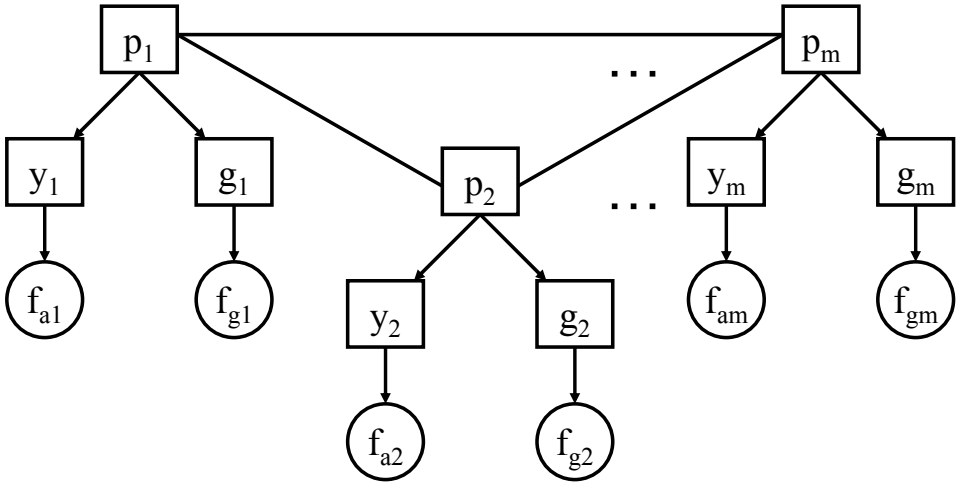


FIGURE 5.6: A graphical model that represents the relationship between a person p having name n, the descriptors of birth year y and gender g, and the associated features f_y and f_g.

5.3.2 First Name Model for Multiple People

When multiple people are in the image, the interactions between the name-person assignments are represented with the graph model shown in Figure 5.6. Our model incorporates the independence assumption that once the identity (first name) of a person p_i is known, the age and gender of this person are independent of the other people in the image. A particular person in the image is p_i, the associated features are f_{ai} and f_{gi}, and the name assigned to person p_i is n_i. We seek to map a set of K first names $\mathbf{N}$ to the set of M people $\mathbf{p}$ in a single image with associated features $\mathbf{f}$ where there are no labeled training faces from which to directly estimate $P(\mathbf{f}|\mathbf{p}=\mathbf{n})$, where $\mathbf{n}$ is a particular assignment of names to people $\mathbf{p}$ in the image.

Using the independence assumptions from the graph model, we write $P(\mathbf{p}=\mathbf{n}|\mathbf{f})$:

$$P(\mathbf{p}=\mathbf{n}|\mathbf{f}) = \frac{P(\mathbf{p}=\mathbf{n})P(\mathbf{f}|\mathbf{p}=\mathbf{n})}{P(\mathbf{f})} \tag{5.4}$$

$$\propto P(\mathbf{p}=\mathbf{n}) \prod_{i=1}^{M} P(f_i|p_i=n_i) \tag{5.5}$$

The maximum likelihood assignment of names to people is the one that maximizes $P(\mathbf{p}=\mathbf{n}|\mathbf{f})$. $P(\mathbf{p}=\mathbf{n})$ is the group prior [48] for a particular set of individuals appearing together in an image. In our case, we assume this term is a non-zero constant only for valid assignments of

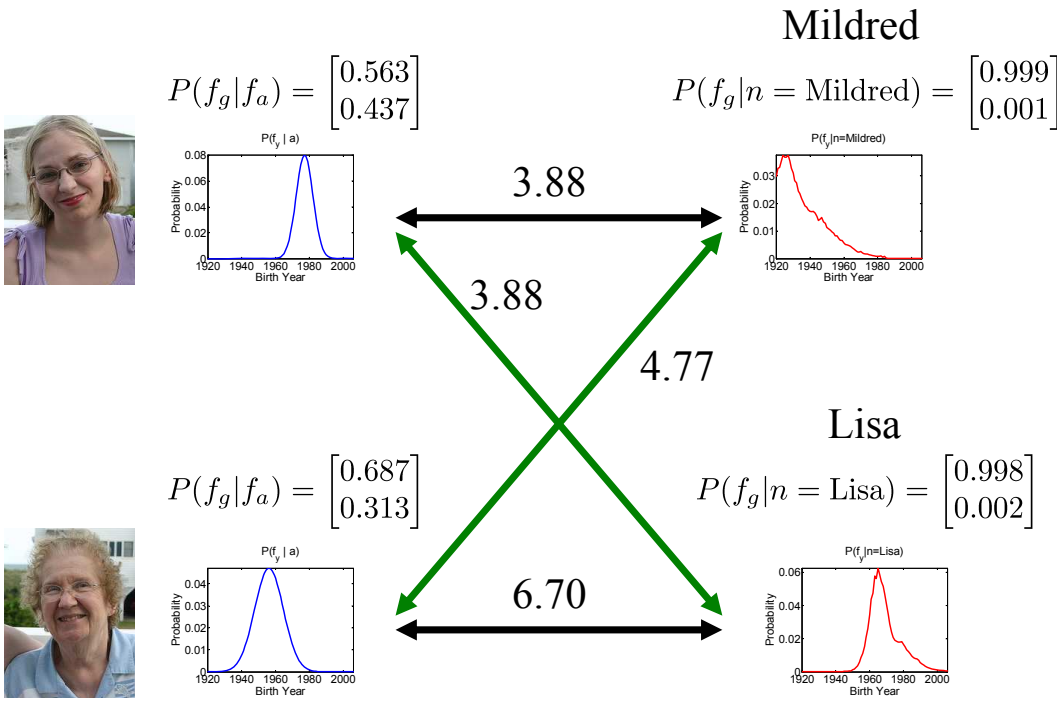


FIGURE 5.7: Assigning names to people can be represented as a bipartite graph. The estimates of gender and birth year given the names Mildred and Lisa as well as the appearance features are shown.The cost of each assignment is shown on each edge, and Munkres algorithm correctly assigns the names to faces (green edges).

names to people. Then, the log likelihood we desire to minimize is:

$$\mathcal{L} = -\log P(\mathbf{p} = \mathbf{n}) - \sum_{i=1}^{M} \log P(f_i|p_i = n_i) \tag{5.6}$$

The term $\log P(\mathbf{p} = \mathbf{n})$ enforces that the name assignments are valid (no more than one person for each name, and no more than one name for each person). Name assignments $\mathbf{p} = \mathbf{n}$ with zero probability incur an infinite penalty. Assuming K first names and M people in the image, there are at most $\max(M, K)!$ possible combinations of names to people to consider. However, the complexity is reduced by recognizing that equation 5.6 exactly describes the classic assignment problem. The assignment problem is represented as a bipartite graph where one set of nodes represents people in the image, and the other set represents first names, as illustrated in Figure 5.7. The cost between each vertex is $-\log(P(f_i|p_i = n_i))$. This problem is solved in $O(\max(M, K)^3)$ using Munkres algorithm [87].

According to our model, age is influenced by both the first name and the age-specific features extracted from the image of the person. Likewise, gender is affected by both the first name and

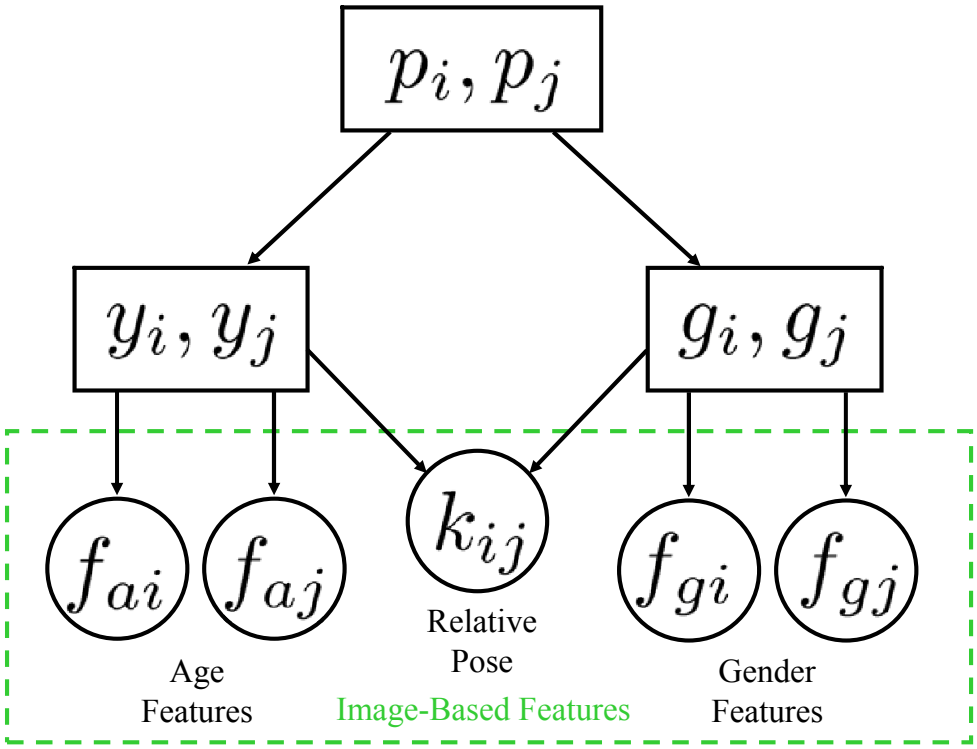


FIGURE 5.8: A graphical model that represents the relationship between a person p with a specific first name, relative pose, the descriptors of birth year y and gender g, and the associated features f_y and f_g.

gender-specific features. Our model is used to select the most likely name to person assignment, and also to refine the image-based estimates of age and gender. In the inference step, we first find the maximum likelihood name assignments $\mathbf{n}^*$ given the initial age and gender estimates, then update the age and gender estimates by finding the marginal distributions over age and gender with our model:

$$P(g|f_g, p = n^*) \propto P(g|p = n^*)\frac{P(g|f_g)}{P(g)} \tag{5.7}$$

A similar calculation is performed for using the model to find the distribution over age $P(y|f_a, p = n)$.

5.3.3 A Model for First Name and Relative Pose

The relative pose between two people in an image is related to the ages and genders of the pair. In Fig. 5.8, we extend the model of Fig. 5.6 by including the relative pose between each pair of people. As relative pose k_{ij} is a pair-wise feature, this model's nodes represent attributes

between pairs of persons i and j, either first name identities (p_i, p_j), ages (a_i, a_j), or genders (g_i, g_j). When more than two people are present in an image, then age, gender, and identity nodes exist for each pair of people in the image, and the relative pose of each pair of people is considered.

The relative pose model represents the joint distribution of age, gender, and identity conditioned on observed image-based age and gender features and observed relative pose. For clarity, we shows nodes of pairs of variables. The graphical model is a correct representation of the joint distribution. However, we assert several independence assumptions that are not directly shown in the model. For example, the joint distribution of genders of two people is not independent of the relative pose of the people, but given the first names and relative poses of a pair of people i and j, the gender of the i^{th} person is independent of the appearance of the j^{th} person, given the gender of the j^{th} person g_j. That is:

$$P(g_i, g_j | f_{gi}, f_{gj}) \propto P(g_i | f_{gi}) P(g_j | f_{gj}) \tag{5.8}$$

With this in mind, we define the conditional probability of the model over the birth year, gender and name of the people in the image conditioned on image-based features and relative pose as the product of unary and pair-wise terms:

$$P(\mathbf{p}, \mathbf{y}, \mathbf{g} | \mathbf{f}, \mathbf{k}) \propto P(\mathbf{p}) \prod_{i=1}^{M} \Psi_i(\mathbf{I}_i) \prod_{i,j=1}^{M} \Phi_{i,j}(\mathbf{I}_i, \mathbf{I}_j) \tag{5.9}$$

The variable $\mathbf{I}_i = \{p_i, y_i, g_i\}$ comprises the set of demographic variables of first name p_i, birth year y_i and gender g_i associated with the i^{th} person in the image. The unary terms Ψ_i, also present in the multiple person model of Fig. 5.6, describe the direct relationships between image appearance and individual attributes of birth year, age and gender as well as the relationship between the attributes and the context provided by first name. The pair-wise terms $\Phi_{i,j}$ describe the relationship between the relative pose of a pair of faces and the distribution of their ages and genders. When the relative pose variable is omitted, the model simplifies to the first name model of Fig. 5.6.

The term $P(\mathbf{p})$ in (5.9) is the group prior. In this model, a pair-wise representation with factors of $\theta(p_i, p_j)$ is used for the group prior. Again, for our purposes, the group prior simply ensures

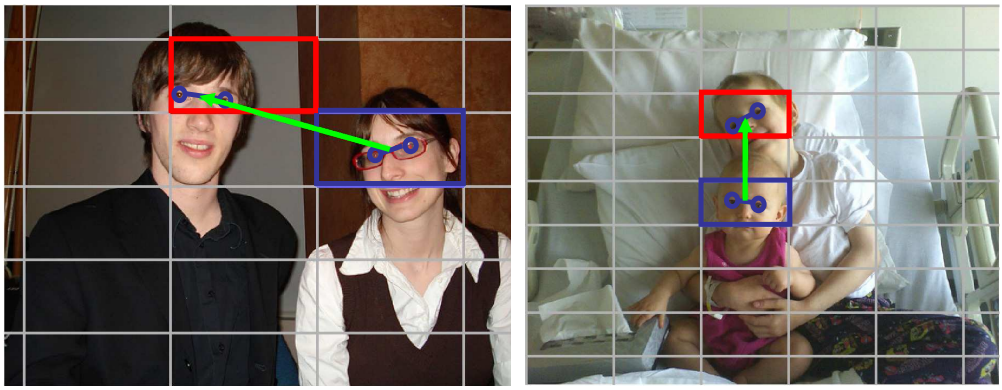


FIGURE 5.9: The relative pose between two faces is quantized into bins whose size is normalized by the average intereye distance of the pair. The quantization is finer in the vertical direction to capture the height differences between the people that provide context for our model.

that no two faces are assigned to the same first name.

$$P(\mathbf{p} = \mathbf{n}) \propto \prod_{i,j=1}^{M} \theta(p_i, p_j) \tag{5.10}$$

The factor $\theta(p_i, p_j)$ is zero when a name is assigned to more than one face.

$$\theta(p_i, p_j) = \begin{cases} 1, & \text{if } p_i \neq p_j \\ 0, & \text{otherwise} \end{cases} \tag{5.11}$$

Unary Terms: The unary terms of the model are factors that capture the relationship between appearance, birth year, and gender as well as first name, as given in (5.2). The m subscript is omitted for clarity.

$$\Psi(\mathbf{I}) = P(p, y, g|f_a, f_g) \tag{5.12}$$

$$= P(p)P(y|p)\frac{P(y|f_a)}{P(y)}P(g|p)\frac{P(g|f_g)}{P(g)} \tag{5.13}$$

Pair-wise Relative Pose Terms: The pair-wise terms of the model are factors that capture the relationship between appearance, birth year, and gender as well as first name.

$$\Phi_{i,j}(\mathbf{I}_i, \mathbf{I}_j) \propto \frac{P(y_i, y_j|k_{ij})}{P(y_i, y_j)}\frac{P(g_i, g_j|k_{ij})}{P(g_i, g_j)} \tag{5.14}$$

k_{ij} represents the quantized relative pose between the faces of two people in the image. The relative pose between two faces is defined as the position of the second face relative to the first. To find the relative pose, the eyes of each of the pair of faces are located with an Active Shape Model [30]. 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 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. Fig. 5.9 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.

The model captures the influence shared between first names, age, gender, and the observed image features and relative pose. Not only can the model be used for finding likely assignments of names to faces, but it can also be used to refine image-based age and gender estimates, as was also the case for the model in Section 5.3.2. After the maximum likelihood name assignments are found, the age and gender estimated are updated by finding the marginal distributions over age and gender with our model. For each person, the model is used to find the maximum a posteriori probability estimate for gender, and the birth year estimate is refined by finding the mean over the posterior birth year distribution.

5.4 Learning Relative Pose Context

The factors in the social context probabilistic models represent empirical distributions and are learned using MLE counts by analyzing training data. As previously mentioned, the first name factors ($P(y|p = n), P(g|p = n), P(y), P(g)$) are learned from the U.S. Social Security baby name database [128] while considering life expectancy [5]. This section deals with learning the parameters for the relative pose factors. The pair-wise relative pose terms capture the relationship between relative pose and the ages and genders of pairs of people. The parameters are learned from training data in two different ways.

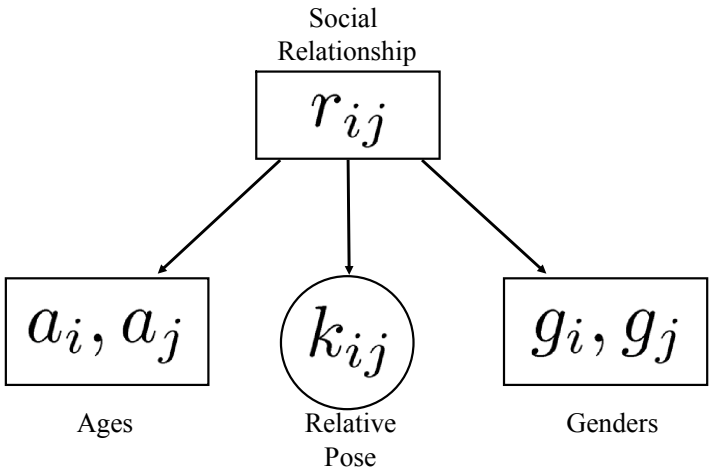


FIGURE 5.10: Social relationship r is used as a latent variable for using publicly available demographic data for learning factors that capture the interactions between relative pose, age, and gender of pairs of people.

5.4.1 Learning From Labeled Images

Given a collection of images where each person's age and gender is indicated, the terms $P(y_i, y_j|k_{ij})$ and $P(y_i, y_j|k_{ij})$ are learned by determining the quantized pose index for each pair of faces across all images, then counting the occurences of pairs of gender and birth year combinations for each pose. Each pair of faces produces two observations, considering each face's position relative the other. In addition, we flip each image left-right to effectively double the training set size. For the birth year factor, soft counts are used when estimating $P(y_i, y_j|k_{ij})$ from the training data. Rather than incrementing only the bin corresponding to the ages of the two individuals, neighboring bins are also incremented. In addition, we assume that the training image could easily have been captured a few years earlier or later, thereby aging each person of the pair equally. In practice, this is accomplished by blurring the accumulator with a Gaussian filter with a width of two years per standard deviation for each individual and six years along the diagonal axis.

5.4.2 Learning From Images and Demographic Data

In an alternate approach, the power of a huge amount of demographic data is used to learn the factors $P(y_i, y_j|k_{ij})$ and $P(y_i, y_j|k_{ij})$ that describe the relationship between relative pose and the ages and genders of pairs of people. In this approach, we require that images be labeled with the social relationship r_{ij} between the people in the image (rather than the age and gender

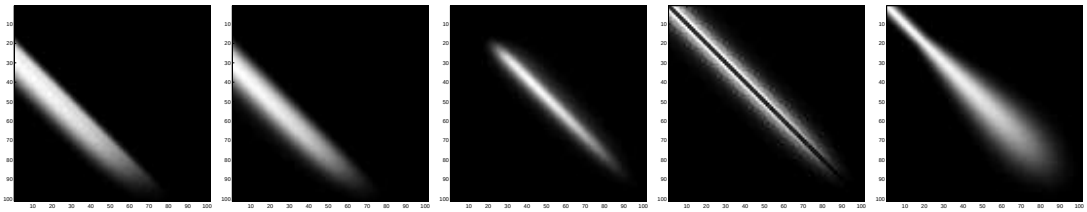


FIGURE 5.11: Each image is a representation of $P(A_1, A_2|R)$, the age of a first person (vertical axis) and a second person (horizontal axis) sharing a social relationship R. The relationships are, from left to right: "mother-child", "father-child", "wife-husband", "siblings" and "friends". Except 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. The joint distribution of ages of siblings is bimodal for biological reasons (twins are rare, accounting for about 3% of births [83]). Husbands are, on average, older than their wives, and it follows that the age gap between father and their children is greater than between mothers and their children.

of each person). Using publicly available data, it is possible to model the distributions between the ages of people involved in different social relationships, as shown in Figure 5.11, using demographic statistics from [5, 14, 24, 39, 83, 128]. By modeling the relationship between the ages, genders, relative pose and social relationship with a graphical model as shown in Fig. 5.10 (a Bayes network in this case), it is then easy to learn the factors $P(y_i, y_j|k_{ij})$ and $P(g_i, g_j|k_{ij})$ using empirical counts and marginalizing over social relationships as follows:

$$P(r, a_i, a_j, g_i, g_j, k) = P(r)P(a_i, a_j|r)P(k|r)P(g_i, g_j|r) \tag{5.15}$$

$$P(a_i, a_j, k) = \sum_r P(r)P(a_i, a_j|r)P(k|r) \tag{5.16}$$

$$P(g_i, g_j, k) = \sum_r P(r)P(g_i, g_j|r)P(k|r) \tag{5.17}$$

The social relationship variable r takes a value from the following set of social relationships:

`Mother-Child`	`Child-Mother`	`Siblings`
`Father-Child`	`Child-Father`	`Friends`
`Husband-Wife`	`Wife-Husband`	`Other`

TABLE 5.1: Social Relationships

In learning $P(g_i, g_j|k_{ij})$, the factor $P(g_i, g_j|r)$, the distribution of genders for each relationship, is easy to estimate from the definitions of the relationship. For example, a child can be a male or female with equal likelihood, but a mother is always female. Publicly available demographic data is used to estimate $P(a_i, a_j|r)$. For the mother-child relationship, Tables 2 and 3 of [83] provides the data necessary to compute model the age difference between a mother and child. Likewise, Table 21 of [83] provides the data for modeling the age difference between father

and child. The age difference distribution between siblings is found in Table 13 of [24] which details the distributions in months between births for women. Two small studies show the joint distribution of ages of husbands and wives, Tables md5 and md6 of [39] and [14]. We model the age distribution between friends as follows: older people have more tolerance to age differences for friends than younger people. The social relationship "Other" is modeled simply as a random selection of two people, drawn according to the distribution of ages. Fig. 5.11 shows a visualization of the joint distribution of ages for five of the relationships.

The terms $P(r)$, the relationship prior and $P(k_{mn}|r)$ are estimated from several family image collections (similar to publicly available [49]). In total, 700 images with multiple people having known social relationships from 12 family image collections are used along with the demographic data to estimate the model parameters for pose. In these collections, the identity of each face is labeled and the social relationship is known for each pair of people.

Fig. 5.12 provides insight into the learned relative pose parameters. Each row shows four images from the test set, with all images in a row having the same quantized relative pose. In many cases, the ages and genders across images within a row are similar.

5.5 Image-Based Gender and Age Classifiers

Our model requires estimates of $P(y|f_a)$, age given age-specific features and $P(g|f_g)$, gender given gender-specific features extracted from an image.

We implemented age and gender classifiers following the examples of [56, 76] and [8, 138]. For age classification, we acquired the image collections from three consumers, and labeled the individuals in each image, for a total of 117 unique individuals. The birth year of each individual is known or estimated by the collection owner. Using the image capture date from the EXIF information and the individual birthdates, the age of each person in each image is computed. This results in an independent training set of 2855 faces with corresponding ground truth ages. Each face is normalized in scale (49×61 pixels) and projected onto a set of Fisherfaces [11] created from an independent set of faces from 31 individuals. The age of a query face is found by normalizing its scale, projecting onto the set of Fisherfaces, and finding the nearest neighbors (we use 25) in the projection space using a Euclidean distance measure. The estimated age of the query face is the median of the ages of these nearest neighbors. Given this estimate for the age, we then model $P(a|f_a)$ as a Gaussian having a mean value of the estimated age, a

FIGURE 5.12: Learning relative pose factors from demographic data. Each row shows four images with the same quantized relative pose, where the first person's eyes are enclosed in a blue box and the second person's quantized face pose is indicated with the red box. The learned factor $P(a_1, a_2|k_{12})$ indicates the joint distribution between the ages, is shown in the fifth column with the first person's age on the y-axis and the second person's age on the x-axis. The origin is in the upper left. The last column shows the learned $P(g_1, g_2|k_{12})$, the joint distribution of genders of the two persons. Row 1: A face positioned well below another face is usually a small child, but sometimes a female-male couple. Rows 2 and 3: When one face is above another and spatially close, it is usually a (taller) male-female couple. Row 4: Two horizontally adjacent faces are usually roughly the same age.

standard deviation of one-third the estimated age (the accuracy of our age classifier decreases with age), and truncated so that ages less than zero have no density. Figure 5.13 shows several age classification results.

Following the example of [138], we implement a face gender classifier using a support vector machine. We reduce the feature dimensionality by first extracting facial features using an Active Shape Model [30]. The ASM locates 82 key points including the eyes, eyebrows, nose, mouth, and face border. Following the method of [48], PCA is further used to reduce the dimensionality to five features. 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. Figure 5.14 shows the gender estimation results for a selection of face images.

FIGURE 5.13: A sampling of our image-based age estimation results. Each row shows a random selection of people for which the age classification result was within a specific range. (Top) Babies and children under the age of five. (Middle) Adults between the ages of 18 and 41. (Bottom) Adults older than 42. The colored bar indicates whether that classification agreed with the human-estimated age for the person, where green indicates agreement.

5.6 Experiment

Tags are often used to indicate objects within an image without providing the spatial location of the objects. For example Flickr and Adobe Albums software both allow users to tag images with keywords. Our experiments address the scenario where images contain multiple people, and are tagged to indicate the first names of the people in the image. Our goal is to disambiguate the tags by assigning names to people based on a single image and to estimate the age and gender of each person. This name-person assignment could be a useful first step for an application that then searches for these same individuals in other images.

We used the following method to collect test sets of names and faces. For Set A, the U.S. baby name database is used to generate random first names. We produce 100 independent pairs of random names. For example, the first three name pairs are: “Jessica and Geraldine”, “Linda and Rosemary”, and “Steven and Luke”. A search is performed on Flickr to find images containing the pairs of people with those first names. The images from the search were painstakingly examined to manually assign names to faces (using captions and other tagged images from the same user’s collection). Most of the images are 500×375 pixels, and contain people with challenging poses and expressions, difficult lighting, sunglasses, and occlusion. We also kept

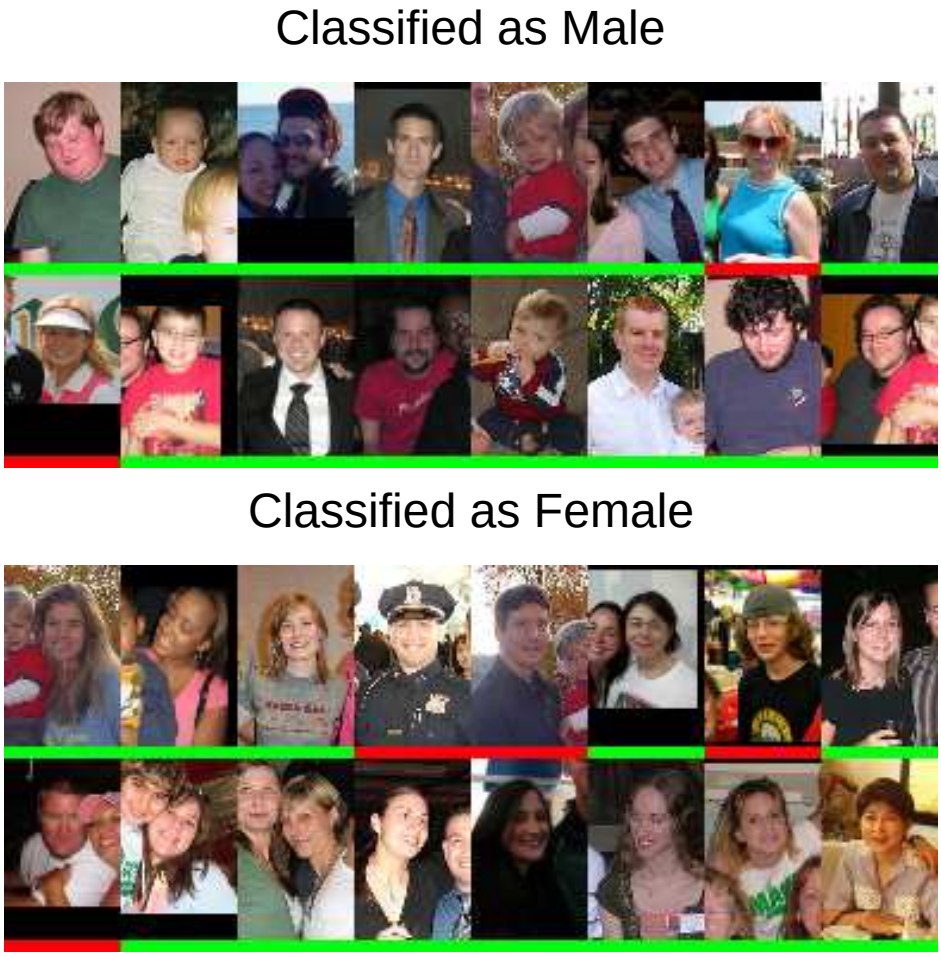


FIGURE 5.14: Gender classification results. (Top) A random selection of people classified as male. (Bottom) A random selection of people classified as female. The colored bar beneath each image is green if the classification is correct.

images where people in addition to the name pair of interest were present, resulting in 34 images with more than 2 people. For most of the name pairs at least one image was located, resulting in a test set of 134 images with 307 people.

In constructing Set B, we selected name pairs that might be difficult for humans to perform the name assignment task. For example, the names “Chris” and “Dana” can each be male or female but each lean towards a specific gender. Also, we used name pairs that have a large disparity in expected birth year, but are perhaps less well known, for example “Tammy” (most popular in the 1960’s) and “Paige” (popular in the past decade). This small but challenging set contains 14 images and the associated first name tags. Set C contains all those images from Sets A and B where all people have a common gender. Name assignment is difficult in this subset since recognizing gender alone is not sufficient to ensure good performance. Table 1 5.2 summarises

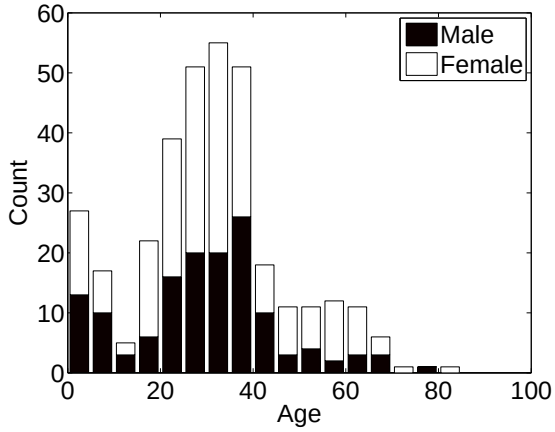


FIGURE 5.15: The distribution of age and gender in the test set.

	Set A	Set B	Set C	Overall
Total images	134	14	48	148
Total people	307	32	105	339
Total males	132	8	26	140
Total females	175	24	79	199
Total children under 10	36	8	12	44
Images with >2 people	31	3	8	34
Uniform gender images	40	8	48	48

TABLE 5.2: A summary of our test sets. Set C is comprised of all images from Sets A and B where the people all have the same gender. The Overall set is the union of sets A and B.

the characteristics of the test images for our experiments, and Fig. 5.15 shows the demographic distribution of the people in the image collection.

For detecting faces, we use a cascade face detector similar to [67]. As our focus is not on face detection, we manually add faces that are missed by our face detector by clicking on the eyes of the missed face. Faces range in size from 12 to 74 pixels between the pupils. We compute image-based estimates of the age and gender of each person using the classifiers described in Section 5.5. Finally, our model (Section 5.3) is used to find the most likely assignment of first names to faces and estimates of age and gender that incorporate evidence from image features and the social context provided by first names and relative pose.

5.6.1 Name Assignment Accuracy

Table 5.3 reports the accuracy of our algorithm at the name assignment task, considering different subsets of the testing set and the model. We show a considerable improvement over random guessing for all subsets of test images. Using the image-based age classifier provides improved

	Set A	Set B	Set C	Overall
Random	43.7%	43.8%	45.7%	43.7%
Age	47.9%	**59.4%**	58.1%	49.0%
Gender	59.3%	56.3%	51.4%	59.0%
Age+Gender	62.2%	56.3%	61.9%	61.7%
Pose	57.0%	43.8%	43.8%	55.8%
PoseA+Age+Gender	**67.4%**	**59.4%**	60.0%	**66.7%**
PoseB+Age+Gender	63.8%	56.3%	**64.8%**	63.1%

TABLE 5.3: Using image-based age and gender classifiers with first name and relative pose as social context improves person recognition in single images. The percentage of correct name assignments is reported. The "Random" row values are expectations rather than an actual experiment. Rows 2-4 show the improved accuracy achieved by using the first name prior along with image-based estimates of age, gender, or both. Row 5 shows the name assignment accuracy using no image features other than the relative poses of pairs of people. Rows 6-7 show that the best performance comes from the integrated model that uses social context from first name priors and relative pose. For the overall set, the results have a statistical margin of error of 3.4%.

name assignment with images of constant gender (Set C), and in the challenging Set B. Using either the image-based age or gender classifiers improves over random first name assignment. By combining age and gender descriptors, greater accuracy is achieved (61.7% overall, versus 43.7% with random assignment).

The best accuracy is achieved by the model incorporating social context from relative pose and first names (Section 5.3.3). In row 6, the model is trained from images labeled with age and gender (leaving-one-out) as described in Section 5.4.1. This model provides a substantial 5.0% improvement (61.7% to 66.7%) over the first name model without pose. In row 7, the model is trained using publicly available demographic data and images labeled only with social relationships. In either case, relative pose provides a benefit over the first name model that omits relative pose. On three of the four subsets, the image-trained pose model achieved the best accuracy. However, it is noteworthy that the relative pose model trained with demographic data achieves the best results on the difficult Set C, images with people of the same gender. These results show that the vast amount publicly available provides a useful tool for learning social context.

Figure 6.9 discusses several image examples, the name assignments, and the age and gender classifications from the image-based classifiers and from the model.

	Age	Gender
	Image→Model	Image→Model
Age	10.0 →**9.33**	→35.4%
Gender	→13.7	28.6% →18.3%
Age+Gender	→9.38	→19.5%
Pose	→9.78	→21.5%
Pose+Age+Gender	→9.91	→**15.3**%

TABLE 5.4: Our model provides improvement over the image-based age and gender classifiers. This table shows the error reduction achieved by estimating age and gender with our model. For the age column, we show that the mean absolute difference between an age estimate and a manually labeled age. For the gender column, the percent is the gender classification error rate. The rows show the error reduction using the image-based age classifier, the image-based gender classifier, both age and gender, relative pose alone, or all the features. The model predicts age even when no image-based age classifier is used, and gender even when no image-based gender classifier is used.

5.6.2 Age and Gender

Our model improves the age and gender estimates over the estimates from the image-based classifiers. For each person image, we manually labeled the age and the gender of the person (without looking an any name information or tags associated with the image). 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. Our model is used to assign first names to people, and then the age and the gender are re-estimated based on the first name assignments by inferring over the model as described in Section 5.3. Both the age estimation and the gender classification are improved through this process, as shown in Table 5.4. For example, using relative pose alone to assign first names to faces (with an overall accuracy of 55.8%), the model can correctly classify gender 78.5% of the time, and has a mean absolute age error of 9.78 years.

In the complete model, the gender classification error is reduced by 46% (28.6% to 15.3%) compared to using only image-based classifiers. The age classification error reduction is modest (10.0 years to 9.9 years error), likely due to the fact that most first names vary only slowly in popularity over time so there is less potential for improvement. Although relative pose is useful for identifying babies, the image-based age classifier generally already accurately identifies them. In all cases, the model reduces the error from the appearance-based classifiers alone. The model can be used to predict age and gender even when image-based appearance classifiers are not used. For example, considering relative pose and first names provides a mean absolute age error of 9.78 years and a gender misclassification of 21.5%, both improvements over the corresponding image-based appearance classifiers (10.0 years and 28.6%).

	Set A	Set B	Set C	Overall
Subject 1	79.2%	81.3%	**65.7**%	79.4%
Subject 2	78.2%	68.8%	61.0%	77.3%
Subject 3	79.5%	43.8%	54.3%	76.1%
Subject 4	69.1%	53.1%	41.9%	67.6%
Human Age+Gender	**80.8**%	**93.8**%	63.8%	**82.0**%

TABLE 5.5: Results for Humans. Four subjects perform the same name assignment task as does our algorithm, and this table reports each subject's accuracy for assigning names to faces. The last row ("Human Age+Gender") reports the results of using our model for name assignment, but using manually labeled values for age and gender rather than image-based classifiers.

5.6.3 Human Performance

It is interesting to compare the results of our algorithm with the accuracy of a human attempting the same task. A user interface was created to allow a human subject to easily assign each tagged name to the person that the subject felt was most plausible. A total of four subjects repeated this exercise for each of the 146 images in the test set. The results of this human experiment are reported in Table 5.5. The values in this table can be compared directly with those for our model, shown in Table 5.3.

Subjects 1 and 2 have the overall best performances and are U.S. born, while subjects 3 and 4 have each lived in the United States for about five years and have lower classification accuracy. This supports our assertion that this image understanding problem requires an understanding of cultural context. Although both subjects 3 and 4 speak fluent English, they have had less time in the U.S. and less time to form this contextual prior, and therefore find the name assignment task more challenging. In fact, by virtue of having a more complete contextual prior, our model outperforms subjects 3 and 4 on the difficult Sets B and C. It is remarkable to note that overall, subject 4 outperformed our social context model by less than 1% (66.7% for our model versus 67.6% for subject 4). Of course, all of the human subjects' visual systems are far more capable than the image-based age and gender classifiers employed by our model. However, the inadequacy of the image-based age and gender classifiers is compensated for in the probabilistic model by using the social context of first name priors and relative pose.

We did an additional experiment to verify our model. Rather than relying on age and gender estimates from the image-based classifiers, we manually labeled each person's age and gender, without any knowledge of the names associated with the image. Then the model is used to produce name assignments using these manually derived estimates for $p(y|f_a)$ and $p(g|f_g)$. The accuracy of this approach is reported in the "Human Age+Gender" row of Table 5.5. This

method produces the highest overall name assignment accuracy compared to the four test subjects, beating the best human subject by 2.6%. This success can be explained by considering that the model has complete domain knowledge regarding first names in the United States, but each human's contextual knowledge of first names is incomplete to some degree. When the model is given a human-level ability to classify gender and age, it is difficult for a human to achieve greater accuracy at the first name assignment problem. From this experiment, we draw several conclusions. First, we expect that improved gender and age predictors will improve the performance of our model. Second, because the performances of the human subjects and the "Human Age+Gender" method are similar, our model is validated and the independence assumptions that we assert are shown to be reasonable.

5.7 Conclusion

In this chapter, a probabilistic model is introduced for integrating social context, appearance, and identity for understanding images of people. Our model integrates the social context elements of the first name prior and the relative pose between people in images. With this model, we infer likely name assignments for images tagged with the first names of the people in a single image. The model learns social context by using publicly available demographic data as well as image data. Further, we show that the model's estimates of age and gender are superior to those from a classifier using solely image-based appearance features.

In a broader scope, this work is a case study emphasizing that images must be interpreted in the context of the culture in which they are captured. We demonstrate learning our social context model from large databases of publicly available demographic data (specifically, data regarding the popularity of first names, and data describing the demographics of people in various social relationships). We believe this chapter represents the first demonstration of using raw demographic statistics as social context to significantly improve a computer vision task. A good understanding of social context provides a strong prior for image understanding.

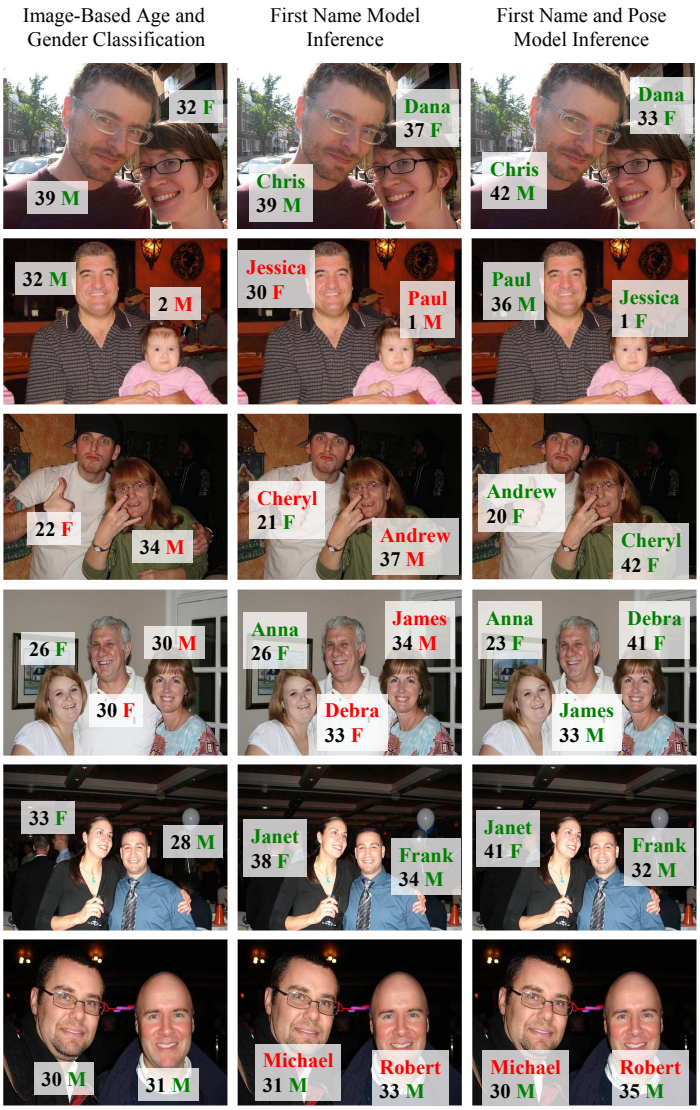


FIGURE 5.16: Our model assigns names to people in images, and improves the performance of gender and age classifiers. The left image in each triplet shows the estimated age and gender directly from our image-based classifiers. (Green indicates correct, and red indicates incorrect. Age classification results are not marked.) The middle image shows the assigned names for each person, and the estimated age and gender from the first name model [50] of Section 5.3.2 and Fig. 5.6. The right image in the triplet shows the assigned names and estimated ages and genders from the complete model with social context from first name and relative pose as described in Section 5.3.3 and Fig. 5.8. Row 1 shows the model makes good guesses for identity even when faces with two gender-ambiguous first names ("Chris" is usually male and "Dana" is usually female). Rows 2-4 show examples where the additional context provided by relative pose allowed the model to correctly identify people and improve age estimates (for Cheryl and Debra). Row 5 illustrates that a rare pose (woman's face higher in the image than the man's) does not confuse the result because the model considers all evidence when assigning names to faces. For images where all people have similar age and gender (as in Row 6), the model assignment is essentially random.

Chapter 6

Jointly Estimating Demographics and Height with a Calibrated Camera

One important problem in computer vision is to provide a demographic description a person from an image. In practice, many of the state-of-the-art methods use only an analysis of the face to estimate the age and gender of a person of interest. We present a model that combines two problems, height estimation and demographic classification, which allows each to serve as context for the other. Our idea is to use a calibrated camera for measuring the height of people in the scene. Height measurement is possible by jointly inferring across anthropometric dimensions, age, and gender using publicly available statistics. The height estimate provides context for recognizing the age and gender of the subject, and likewise the age and gender conditions the distribution of the anthropometric features useful for estimating height.

The performance of our method is explored on a new database of 127 people captured with a calibrated camera with recorded height, age, and gender. We show that estimating height leads to improvements in age and gender classification, and vice versa. To the best of our knowledge, our model produces the most accurate automatic height estimates reported, with the error having a standard deviation of 26.7 mm.

The goal of this chapter is to describe a person's height and demographics from an image. In computer vision research, algorithms exist to identify the age and the gender of people. Broadly speaking, these algorithms build statistical models for the image appearance of a person for different demographic categories, and these models are employed to categorize the image of a previously unseen face. With few exceptions, demographic recognition is performed solely

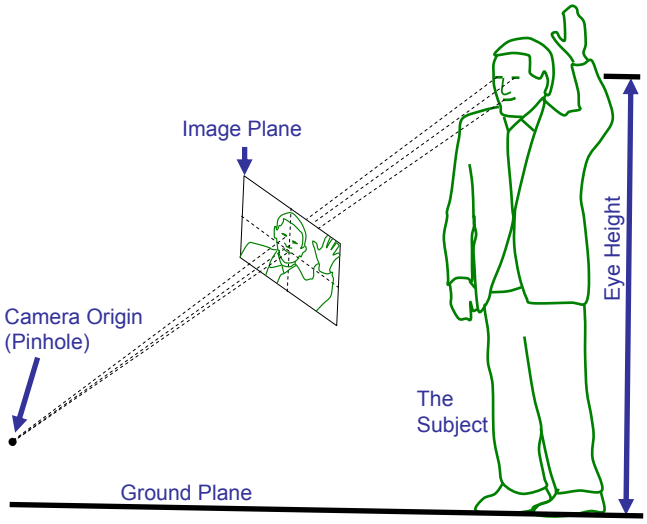


FIGURE 6.1: Our approach for measuring human height with a calibrated camera. Calibration provides the relationship between the image and world coordinate systems. Facial key points fall on rays passing through the camera center and corresponding images of the points. Anthropometric data, conditioned by the estimated age and gender, provides the distribution on distances between key points so distance from subject to camera, height, age, and gender are inferred.

based on facial appearance. In practice, however, facial appearance does not provide enough information to solve this problem with the desired level of accuracy.

Similarly, several researchers have investigated the problem of estimating the height of a standing or walking human. In some cases, the problem has been addressed solely as a metrology problem, using similar techniques than can be applied for measuring any other vertical object.

The goal of our chapter is to unite these two sub-problems (height measurement and demographic estimation) into a common framework employing a probabilistic model to allow evidence gathered for each sub-problem to reduce the uncertainty about the other. Our approach is to combine facial appearance with height estimation to improve our understanding of images of people. To this end, we exploit the large volume of anthropometric measurements gathered by the medical and health communities.

6.0.1 Related Work

A large amount of research is directed at 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. Most age and gender classification algorithms construct feature vectors solely from the face region [8, 56, 58, 63]. In fact, the vast majority

of classification work related to images of people treats each face as an independent problem and relies solely on information gleaned from images from which classifiers are constructed. However, there are some notable exceptions where information external to the image is used as context for classification. In [13], names from news 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 [50], the popularity trends of first names provide context in conjunction with facial appearance to infer age and gender.

Regarding height estimation, several researchers either estimate height, or use broad height distributions with pedestrian detection to understand scenes. The position of people in an image provides clues about the geometry of the scene. As shown in [81], 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). In [70], the problem is reversed, and the height of a person with visible feet and head is estimated from a calibrated camera. Criminisi *et al*. [33], Hoiem *et al*. [66], and Lalonde *et al*. [74] describe the measurement of various objects (including people) rooted on the ground plane. However, all of these papers require that the intersection of the object (i.e. the feet) and the floor be visible. Multiple cameras are employed in [7], turning the problem into an application of shape-from-motion. Our method relies on anthropometric face measurements and requires instead that the face be visible.

Our work essentially uses information from the fields of anthropology and medicine as context for demographic inference in computer vision. In anthropology, the relationships between various body measurements has been studied and exploited to estimate the height of an individual from a single recovered bone [40]. Perhaps the closest work on human height measurement from images is BenAbdelkader and Yacoob [12] where anthropometric data is used in combination with manually identified key points and apriori knowledge of age and gender. We build on this work by automatically locating facial anthropometric features and introducing a model that naturally incorporates the uncertainty over gender and age. As a result, gender, age and height can also be inferred from our model.

Our contributions are the following: We propose a model for measuring the height of a person while jointly estimating age, gender and facial feature points, based on a calibrated camera and anthropometric data. We introduce the idea of combining height estimation with appearance features for demographic recognition, and show that estimating height improves the recognition of

demographic quantities. Further, by performing inference over age, gender, and height simultaneously with our model, we improve the accuracy of height estimation. Finally, we demonstrate the effectiveness of our model on a test set of 127 individuals to achieve height estimates with good accuracy.

In Section 2, we introduce human height estimation with a calibrated camera. In Section 3, we describe data related to anthropometric features. Section 4 describes our model of the relationship between height, gender and age from anthropomorphic data. Finally, in Section 5 we describe experiments that demonstrate the effectiveness of our approach.

6.1 Calibrated Camera Height Estimation

As is well known, a camera can be modeled as a projective pinhole [65] to map world points $\mathbf{X}$ to image points $\mathbf{x}$ according to the following relationship:

$$\mathbf{x} \equiv \mathtt{P}\mathbf{X} \tag{6.1}$$

$$\equiv \begin{bmatrix} \mathtt{A} & \mathtt{b} \end{bmatrix} \mathbf{X} \tag{6.2}$$

where the calibration matrix $\mathtt{P}$ is composed of internal camera parameters $\mathtt{K}$ and extrinsic parameters including a coordinate rotation matrix $\mathtt{R}$, and translation $\mathbf{t}$ as follows: $\mathtt{P} = \mathtt{K} \begin{bmatrix} \mathtt{R} & \mathbf{t} \end{bmatrix}$. In the form shown in (6.2), the 3×3 matrix $\mathtt{A} = \mathtt{KR}$, and the 3×1 matrix $\mathtt{b} = \mathtt{K}\mathbf{t}$. The matrix $\mathtt{P}$ essentially captures the relationship between image and scene points, and allows one to extract metric information from image coordinates. Each point in the image corresponds with a world line passing through the camera center.

6.1.1 Camera Calibration

We perform camera calibration using a checkerboard target according to the method of [146], and shown in Figure 6.2. The checkerboard defines the world coordinate system. As such, we ensure that for one image, the target is held perpendicular to the ground. Consequently, the world coordinate system axes are aligned with the physical ground plane (the y-axis is perpendicular to the ground plane, and the x- and z-axes are parallel to the ground plane). In addition, for this image, the distance h_y from the coordinate origin $\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$ is measured by hand, as shown in Figure 6.2. The floor has the equation $y = -h_y$ in the world coordinate frame.

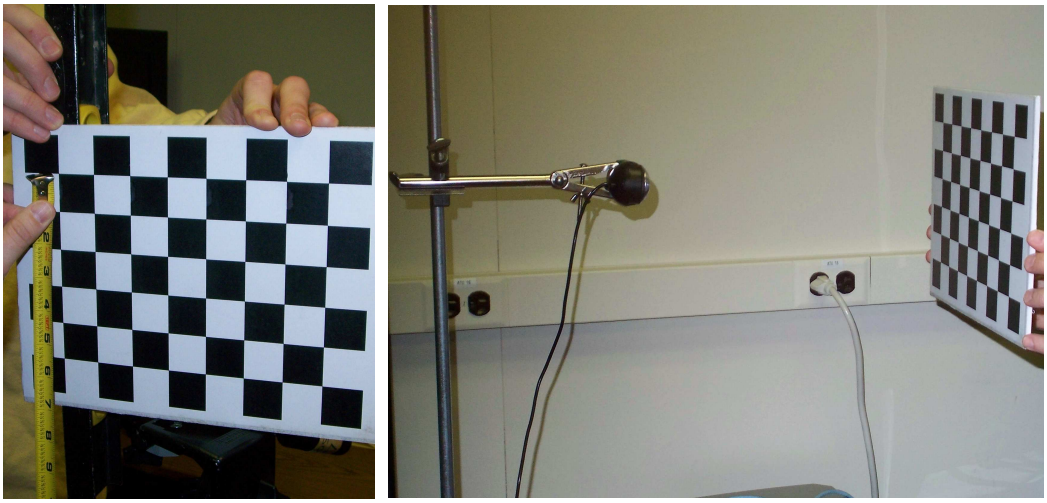


FIGURE 6.2: **Left**: During calibration, for one image a level is used to position the calibration target to be perpendicular with the floor. The distance between the floor and the world coordinate system origin is measured by hand. **Right:** Our camera is a standard web-camera with VGA resolution.

6.1.2 Estimating Subject Distance and Height

Our key idea is illustrated by Figure 7.1: Multiple feature points on a face image corresponding to pairwise anthropometric features define multiple rays in the world. The distribution of possible distances between the camera and the subject is functionally related to the distribution of the size of these anthropometric features. As the uncertainty in the anthropometric feature distribution is reduced (e.g. by concluding that the subject is an adult male), a corresponding reduction in the uncertainty of the distance to the camera is achieved. Furthermore, because the camera is calibrated, an improvement in our confidence about the distance to the subject is directly related to improvements in the determination of the height above the ground plane of each facial feature point.

Estimating Subject Distance: We consider pairwise anthropometric features, defined as the distance between two feature points on the human body. In world coordinates, the pairwise anthropometric feature F is described by a Gaussian distribution $N(\mu_F, \sigma_F^2)$ over a measurement metric. Each feature F has a corresponding pair of image points $\mathbf{f} = \begin{Bmatrix} \mathbf{x}_i & \mathbf{x}_j \end{Bmatrix}$.

A world line $\mathbf{L}$ passing through a particular image feature point $\mathbf{x}_i$ has the equation:

$$\mathbf{L}_i = \mathbf{\Omega} + t\omega_i \tag{6.3}$$

where the camera center is $\Omega = \mathtt{A}^{-1}\mathtt{b}$ and the vector pointing from the camera center to the feature point $\mathbf{x}_i$ is ω:

$$\omega_i = \mathtt{A}^{-1}\mathbf{x}_i \tag{6.4}$$

The angle ϕ between two feature lines $\mathbf{L}_i$ and $\mathbf{L}_j$ is:

$$\phi = \cos^{-1}\left(\frac{\omega_i^T \omega_j}{|\omega_i||\omega_j|}\right) \tag{6.5}$$

and the distance d in world coordinates from the camera center Ω to the midpoint of two feature points on the human body having separation distance d_F is:

$$d = d_F \frac{1}{2\tan(\phi/2)} \tag{6.6}$$

The distribution of the distance d is represented as a Gaussian $N(\mu_d, \sigma_d^2)$ where the parameters are found by considering that (6.6) is a linear function of random variable F. Consequently, $\mu_d = \mu_F \frac{1}{2\tan(\theta/2)}$ and $\sigma_d = \sigma_F \frac{1}{2\tan(\theta/2)}$.

In summary, our knowledge about the distributions of pairwise anthropometric features is exploited to estimate the distance between the subject and the calibrated camera.

Estimating Subject Height: From a subject to camera distance estimate d_i, the feature point can be approximately located (assuming the pair of feature points is parallel to the image plane) in the world coordinate frame as:

$$\hat{X}_i = \Omega + d_i \frac{\omega_i}{|\omega_i|} \tag{6.7}$$

Because our world coordinate frame is axis-aligned with the physical world (the xz-plane is parallel with the ground), the height of a point h_i above the ground is simply:

$$h_i = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \hat{X}_i + h_y \tag{6.8}$$

The estimate for the subject's stature is based on the pairwise anthropometric feature of the eye centers F_e. The stature of a person is the height of the eyes above the ground, plus the distance from the eyes to the top of the head $F_{v,en}$, as reported in [44]. Note that this dimension $F_{v,en}$ has a distribution over gender and age and in practice, the expected value of this distribution is

used.

$$h = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \hat{X}_i + h_y + F_{v,en} \tag{6.9}$$

As with distance, the distribution of height h is represented with a Gaussian, where the parameters are derived by considering h as a function of the random distribution over distance d.

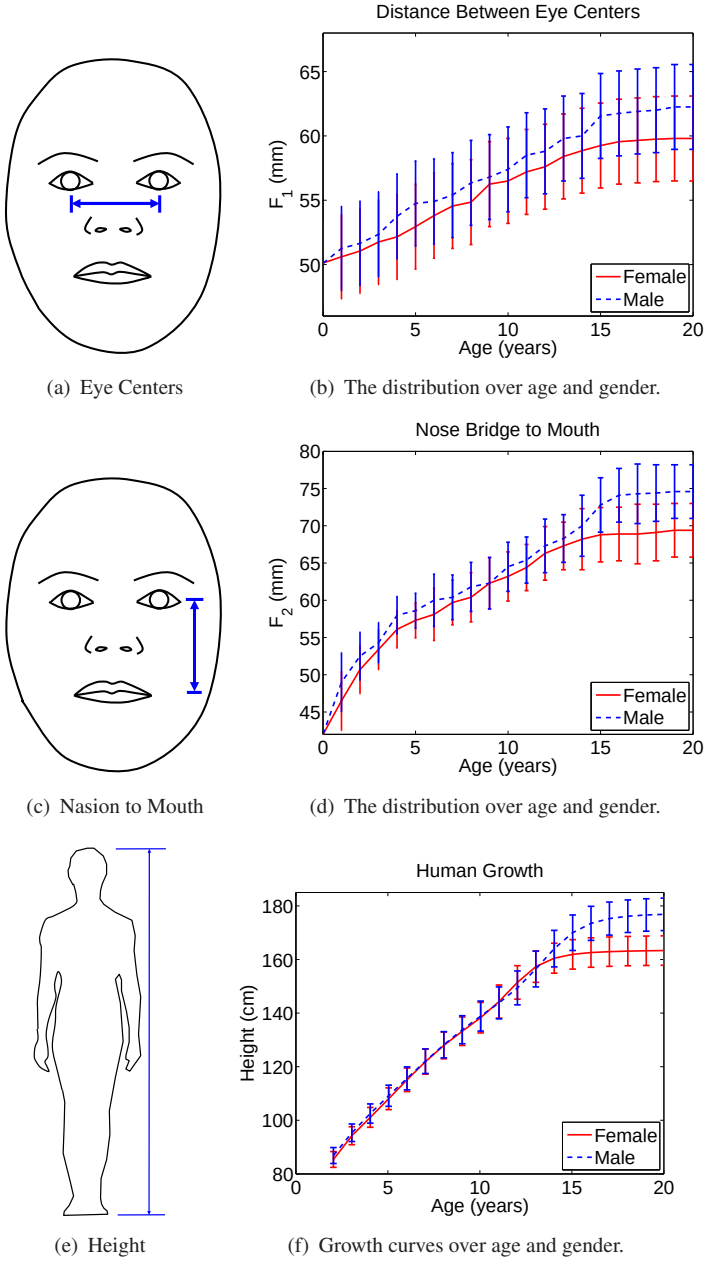
6.2 Age, Gender, and Anthropomorphic Data

There exists a great amount of data describing the distribution of measurements of the human body [44, 60, 92]. Our goal is to use pairwise anthropometric features to infer subject to camera distance, height, age and gender. Ideal anthropometric features are those that markedly change in size with age and gender. We have the additional practical requirement that the corresponding image of each feature point can be reliably located in the image automatically with an Active Shape Model [30].

In this chapter, we reason with two pairwise anthropometric features, illustrated in Figure 6.3. The size distributions as functions of age and gender for each of these pairwise anthropometric features is derived by smoothing data from [44]. The first feature F_1 is the distance between eye centers, and the second F_2 is the distance between the mouth and the nasion (i.e. the intersection of the nose and forehead). Our automatic detection of the associated feature points on several images is shown in Figure 8.3.

6.3 An Anthropometric and Demographic Model

We would like to represent the relationships between a person's age, gender, height and appearance in the image. Of course, our degree of uncertainty about one attribute affects our belief about others. For example, if we are confident that a subject is tall (e.g. 190 cm), then it is more likely that the subject is an adult male than an adult female. However, it is intractable to learn the relationship between all quantities simultaneously. Our model incorporates conditional independence assumptions to make inference tractable and allows inference over all quantities in a unified manner.



(a) Eye Centers

(b) The distribution over age and gender.

(c) Nasion to Mouth

(d) The distribution over age and gender.

(e) Height

(f) Growth curves over age and gender.

FIGURE 6.3: The two pairwise anthropometric features we use in this work are the distance between eye centers (Top), and the distance between the mouth and nasion (the point between the eyes where the nose bridge meets the frontal bone of the skull) (Middle), which have known distributions with respect to age and gender. The relationship between gender, age, and height is also shown (Bottom). Error bars represent one standard deviation.

Figure 6.5 shows a graphical representation of our model. We represent the demographic and anthropometric quantities as random variables in the model. Each subject has an age A, gender G, height H, and distance from the camera D. The true value of the subject's i^{th} pairwise anthropomorphic feature is denoted by the variable F_i and the set of all such features is $\mathbf{F}$. Observed evidence includes a set of image points for each pairwise anthropometric feature $\mathbf{f}$,

FIGURE 6.4: Example images with automatically recovered key points corresponding to two pairwise anthropometric features. The eye center distance is related to the distance between the circles, and the mouth to nasion feature points are marked with the symbol '+'.

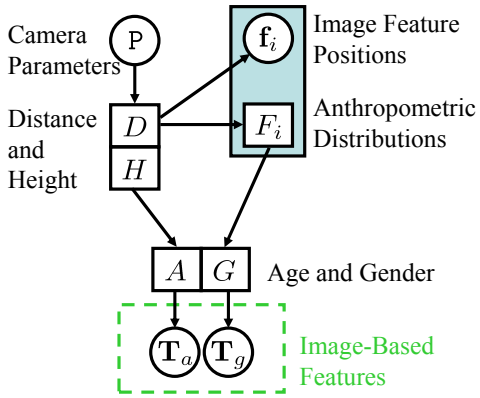


FIGURE 6.5: Our graphical model to infer over age A, gender G, height H, and camera to subject distance D, based on the evidence that includes the camera parameters $\mathtt{P}$, the extracted feature points $\mathbf{f}_i$, the anthropometric feature distributions F_i and the appearance features $\mathbf{T}_a$ and $\mathbf{T}_g$ related to age and gender respectively. Hidden variables are squares, with adjacent squares representing joint variables, and observed variables are circles.

the camera calibration parameters $\mathtt{P}$, and appearance features extracted from the pixel values of the face region corresponding the age $\mathbf{T}_a$ and gender $\mathbf{T}_g$. Our model includes simplifying conditional independence assumptions. For example, we assume that once age and gender are known, the facial appearance is independent of the height of the subject. Further, once the subject height and pairwise anthropometric measurements are known, the calibration parameters provide no further insight regarding the subject's demographic information. The structure of the Bayes Network is selected to exploit known relationships documented with publicly available statistics as well as known relationships from perspective geometry.

The model represents $P(A, G, H, \mathbf{F}|\mathtt{P}, \mathbf{f}, \mathbf{T}_a, \mathbf{T}_g)$ as a product of conditional probability terms. Gaussians are used to represent the distributions over variables related to distance (D, H and

$\mathbf{F}$). Gender G is a binary variable $G \in \{\texttt{male}, \texttt{female}\}$. Age A is a discrete variable with a set of 125 possible states corresponding to the ages 0 to 124 years. In the following sections, we describe the terms of our model and inference with the model.

6.3.1 Estimating Age and Gender from Appearance

Our model employs appearance-based age and gender classifiers. These content-based classifiers provide probability estimates $P(G|\mathbf{T}_g)$ and $P(A|\mathbf{T}_a)$ that the face has a particular gender and age category, given the corresponding visual appearance features.

Our gender and age classifiers were motivated by the works of [56, 63] where a low dimension manifold for the age data. An independent set of 4550 faces is used for training. The age and gender of each person was labeled manually. To establish age ground truth, 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. 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 each pair of age categories, resulting in 21 columns for $\mathbf{W}_a$. A similar approach is used to learn the gender subspace $\mathbf{W}_g$. A set of seven projections is found by learning a single projection that maximizes gender separability for each age range.

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

$$d_{ij} = (\mathbf{T}_i - \mathbf{T}_j)\mathbf{W}\mathbf{W}^T(\mathbf{T}_i - \mathbf{T}_j)^T \tag{6.10}$$

For classification for both age and gender, the nearest N training samples (we use $N = 101$) 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 $P(A|\mathbf{T}_a)$ and $P(G|\mathbf{T}_g)$ 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 the discriminative projections are modified.

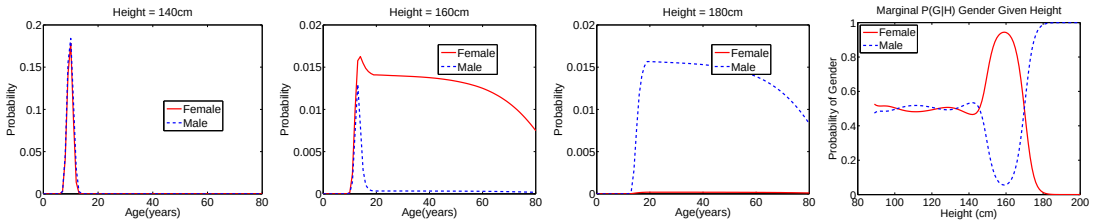


FIGURE 6.6: Illustrations of $P(A, G|H = h)$, the joint distributions over age and gender given height, for several different heights h. **Top Left:** When $h = 140$ cm, the subject age distribution is centered at 10 years with nearly equal likelihood of each gender. **Top Right:** There are two reasonable explanations when height is $h = 160$ cm. Either the subject is an adult female, or an adolescent male in the process of "growing through" that height. **Bottom Left:** A person with a height $h = 180$ cm is most likely a male. **Bottom Right:** The marginal distribution of gender given height. Note the peaks at heights common for adult women and men.

6.3.2 Anthropometrics from Age and Gender

Our model requires the term $P(A = a, G = g|F_i)$, the conditional distribution of age and gender given a particular pairwise anthropometric feature F_i. This term is provided by the statistical data of [44], illustrated in Figure 6.3 for the two anthropometric features we consider.

6.3.3 Distance and Height

The relationship between the camera parameters $\mathtt{P}$, the pairwise demographic features $\mathbf{F}$, the corresponding features $\mathbf{f}_i$ in the image, and distance to the subject D is a deterministic function of random variables, described with Equations (6.3)-(6.6). Therefore, the term $P(D|F_i, \mathbf{f}_i, \mathtt{p})$ is simply a function of a random variable, where the distribution of F_i is related to the the distribution of D. Likewise, the term $P(H|D)$ is also a deterministic function of the random variable distance D (6.7)-(6.9).

6.3.4 Height, Age, and Gender

Our model requires the term $P(A, G|H = h)$, the conditional distribution of age and gender given height. This term is provided by the statistical data of [92] and is illustrated in Figure 6.6. The conditional probability of age and gender given height is found with $P(A, G|H) \propto P(H|A, G)P(A)P(G)$, with the simplifying assumption that age and gender are independent. The gender prior $P(G)$ is assumed to be equal for each gender ($P(G = \mathtt{male}) = 0.5$), and the prior for age $P(A)$ is based on life expectancy from a standard actuarial table [5].

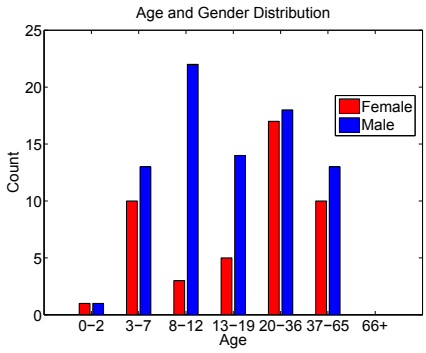


FIGURE 6.7: The distribution of the 127 subjects used in our study.

We make several observations. First, the conditional distribution $P(A, G|H)$ is not well-modeled with a Gaussian distribution because of the rapid growth in the adolescent years, justifying our decision to represent age as a discrete variable. Second, we note that for adults aged 20 or greater, 170 cm represents the optimal decision boundary to classify gender when height is the only available information. Finally, we mention that our model does not consider the phenomena of stature loss among the elderly, but this effect could be added if the relevant statistical data are available.

6.3.5 Inference with Expectation Maximization

We perform inference on our model to consider all the evidence from an image of a subject captured with the calibrated camera, and find the distribution over age, gender, height and distance to the camera. Final classifications are based on the maximum likelihood aposterior distributions for age $\hat{a}$, gender $\hat{g}$, height $\hat{h}$, and distance $\hat{d}$ from the camera. For each variable, our final estimate is the assignment that maximizes its marginal distribution obtained by marginalizing over all other variables.

For computational efficiency, we do not perform exact inference over the entire model. Instead, similar to [122], we use Expectation Maximization to simplify inference. In the E-step, we fix the distribution over F_i as a unidimensional Gaussian and perform inference on the model. In the M-step, the distribution over each anthropometric feature F_i is updated using the winner-take-all variant of EM [93] based on the most likely estimate of age a^* and gender g^* as $P(F_i|A = a^*, G = g^*)$. In our case, the winner-take-all variant has the advantage that, in inference, each anthropometric distribution remains a Gaussian. After convergence, the most likely assignment of each variable is found.

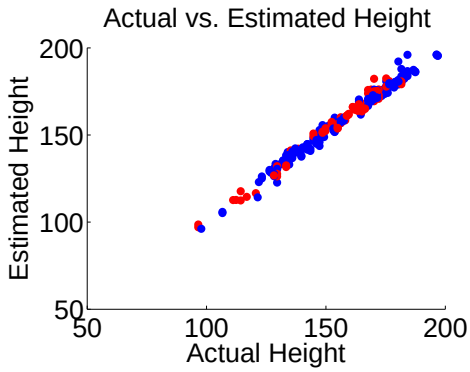


FIGURE 6.8: A scatter plot of the estimated and actual height (cm) of subjects in our study.

	Height		Distance		Age		Gender
	MAD	STD	MAD	STD	MAD	STD	Error
Height	30.5	40.9	182	201	-	-	-
Model+$\mathbf{T}_a, \mathbf{T}_g$	-	-	-	-	8.5	12.3	32.8%
Model+P, $\mathbf{f}_i$	**24.1**	**26.7**	142	**167**	7.0	10.6	35.3%
Full Model	**24.1**	**26.7**	**136**	171	**5.4**	**9.7**	**28.1%**

TABLE 6.1: By reasoning about gender, age and height with our full model we achieve the best overall results for predicting age and gender. Errors (mean absolute and standard deviation) are shown for height and distance. Age errors are in years, and gender classification error rate is shown. Results are shown for height alone (no modeling of age or gender), using the model but observing only appearance features, using the model but observing only height (no appearance), and using the full model.

6.4 Experiments

Our model was tested on images of 127 subjects ranging in age from 2 to 56 with a total of 81 male and 46 female subjects. To sample from a wide variety of demographics, subjects were recruited in several different venues (a science museum, a research laboratory, and an educational institution) on four different occasions. The gender and age distribution of subjects is reported in Figure 6.7. Most subjects are Caucasian, but a wide variety of ethnicities participated. Each subject reported his or her age (binned into one of 14 bins) and gender, and a stadiometer was used to measure each subject's height. Subjects were photographed looking toward the camera. The camera height is about 160 cm off the ground, but this varied at each session. Two pieces of tape were placed on the floor at different distances from the camera, one at a near position (ranging from 0.91 m to 1.63 m) and one at a far position (ranging from 1.80 m to 2.69 m). Each subject was photographed at the two distances marked by the tape. The camera has VGA resolution (480×640 pixels). The entire procedure requires about five minutes for each subject. A total of 237 images are used in our experiments (two images for most subjects; 17 subjects have only one image).

	Height		Distance		Age		Gender
	MAD	STD	MAD	STD	MAD	STD	Error
Multi-frame	**22.4**	**22.1**	-	-	6.2	9.8	**24.5**%

TABLE 6.2: Additional accuracy improvements are achieved by using evidence from multiple images. Compare with the last row of Table 1.

For detecting faces, we use a commercial package that implements a cascade face detector similar to [67]. After face detection, an active shape model [30] is applied to recognize key points on the face, as illustrated in Figure 8.3. Finally, for each subject image, inference is performed with our model in Figure 6.5 to obtain maximum likelihood aposterior estimates for age $\hat{a}$, gender $\hat{g}$, height $\hat{h}$, and distance $\hat{d}$ from the camera.

6.4.1 Height and Distance Accuracy

Table 6.1 reports the accuracy of the model on our test set for height, distance to the subject, age, and gender. We compare height estimation with the baseline approach where age and gender are not in the model, and the anthropometric distributions are from the entire population, marginalizing over age and gender. Overall, the complete model estimates human height with an accuracy of 26.7 mm in standard deviation, reducing the error of the baseline approach by 34.7% (from 40.9 mm). Figure 6.8 shows a scatter plot of the true and estimated statures of the subjects.

This result is believed to be the most accurate automatic result achieved for this task on a large dataset. In [12], estimation error of about 50 mm in standard deviation is reported on a test set of 27 adults, where the model has full knowledge of gender and feature points are manually labeled. In [32], a reference length from the scene is required, and the result on a single subject is within 2 cm. Finally, in [70], height is estimated by a calibrated camera detecting the full silhouette of the subject. On three subjects, this achieves an estimation error with standard deviation of 43 mm.

We estimate the distance between the subject and the camera with an accuracy of 171 mm in standard deviation. This represents the distribution of the distance estimates differences from the two tape marks on the floor that each subject was asked to stand on. In reporting this result, it is noted that the distance to the subject is somewhat variable as each subject's interpretation of "standing on the tape" varied. Therefore, we expect that our reported results represents an upper (i.e. pessimistic) bound on the achievable distance accuracy.

	Stature		Distance		Age		Gender
	MAD	STD	MAD	STD	MAD	STD	Error
Children(0-16)	22.4	22.3	106	116	**0.5**	**0.9**	29.6%
Adults(17+)	22.3	22.9	120	130	11.6	11.9	**19.6%**

TABLE 6.3: Age classification is an easier problem for children, and gender classification is easier for adults. Height estimation performs well across age. These results include using evidence from two images of the same subject, when available.

6.4.2 Combining Multiple Observations

Evidence from multiple observations is combined to estimate the age, gender, and height of a person using a Naïve Bayes model with an assumed uniform prior over the variable in question. For example, when estimating height from multiple images:

$$P(H|\mathbf{e}_1, \ldots, \mathbf{e}_N) = \prod_{n=1}^{N} P(H|\mathbf{e}_n) \tag{6.11}$$

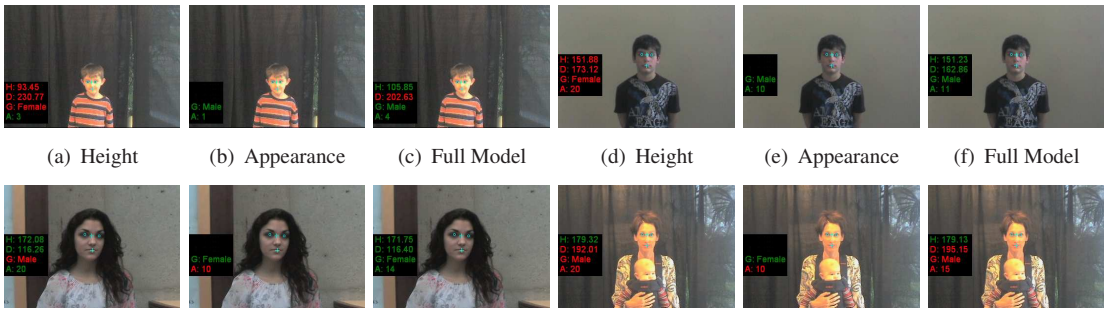
where $\mathbf{e}_n$ represents all the available evidence associated with the $n^{\mathtt{th}}$ image capture.

Table 6.2 reports the result of consolidating evidence from multiple frames (both the near and far image captures) for each subject. Overall, more accurate height estimates and gender classifications are achieved, but the age estimation suffered.

6.4.3 Gender and Age Accuracy

By using our model to infer gender and age using both appearance and height, we achieve better accuracy than using either one alone, as reported in Table 6.1. Our appearance classifier achieves 67.2% gender accuracy by itself. This is lower than the results reported for this task using facial appearance (e.g. [8]), but our test set includes a large number of children who have yet to develop gender-specific facial features. Combining height with appearance by our model improves the gender classification accuracy to 71.9%.

Each subject self-reported his or her age as belonging to one of 14 age bins. Using our model, we find the most likely aposterior age $\hat{a}$, and compare this with the ground truth age bin for the subject. When $\hat{a}$ falls within the bounds of the age bin, the age error is zero, otherwise the age error is the number of years between the estimated age $\hat{a}$ and the closest bound on the true age bin. Again, by inferring age with combined appearance and height features, we achieve better age estimation than using either feature type alone.



(a) Height (b) Appearance (c) Full Model (d) Height (e) Appearance (f) Full Model

(g) Height (h) Appearance (i) Full Model (j) Height (k) Appearance (l) Full Model

FIGURE 6.9: Height, age, and gender classification improve through our model that reasons over variables related to appearance, height, demographics and pairwise anthropometric features. In each group of images, the model outputs are shown when height is observed (no appearance features), appearance is considered (height is not estimated), and the full model is used. Accurate results are shown in green text and poor results are in red text. The facial appearance in (b) allows the mistaken gender from height alone (a) to be corrected in the full model (c). In (d), the subject's height is similar to an adult woman, but appearance recognized the subject as a young male (e), and the full model finds the most probable explanation is that the subject is an adolescent male (f). The incorrect age classification from appearance alone (h) is corrected by height estimation in (g) to produce the reasonable estimates in (i). A failure is shown in (j)-(l). The subject is a tall female, and the correct gender from appearance (k) is not strong enough to override the fact that few females are 179 cm in height from (j), and in the final result (l), the demographic classification is worse than from appearance only (k). Best viewed electronically.

More insight is gleaned by examining the performance on children (ages 0-16) and adults (17+). Table 6.3 shows that age is easier to estimate for children, and gender classification is more accurate in adults. This result is explained by considering our pairwise anthropometric features, as shown in Figure 6.3. For age estimation, the gradient of each feature with respect to age is greatest during childhood. However, the greatest separation between the genders for the distributions for any of the anthropometric features given age occurs when adulthood is reached.

Figure 6.9 discusses the height, age, and gender estimates for several images from our dataset.

6.5 Conclusion

In this chapter, we introduce a model to unify inference over demographic quantities and anthropometric features using a calibrated camera. Instead of considering demographic classification and height estimation as separate problems to be solved independently, our model merges these problems and allows influence to flow throughout the variables.

We provide evidence of the effectiveness of our model by testing on images from 127 subjects spanning a wide age range to achieve an automatic height estimation error of 26.7 mm in standard deviation. We show that when height provides context and is considered along with facial appearance, the age and gender estimates improve versus using appearance alone. Likewise, height estimation improves with our model which reasons about age and gender as hidden variables. Our model is extensible in that additional pairwise demographic features can easily be added, assuming the corresponding feature points can be located in the image.

Chapter 7

Clothing Cosegmentation for Recognizing People

In collections of consumer images, it is a worthwhile task to label each face with its proper identity. In this application, the collection generally comprises hundreds or thousands of images, and the people in the collection often appear many times. The remaining chapters of this dissertation are directed at understanding the role of context in recognizing people in consumer image collections. In this chapter, we examine the role of clothing as context. In Chapter 8 we explore the role of the group prior, a learned prior over specific groups of people in the collection. In Chapter 9, multiple contextual features are integrated into a single model. In all of these chapters, we address the problem in the scenario where some portion of the image collection faces are labeled, and the model is used to infer the identity of the remaining faces. The question of choosing the most informative subset of faces to label is addressed in Appendix A.

To overcome the limitations of face recognition in consumer images, features other than faces need to be considered. Reseachers have verified that clothing provides information about the identity of the individual. To extract features from the clothing, the clothing region first must be localized or segmented in the image. At the same time, given multiple images of the same person wearing the same clothing, we expect to improve the effectiveness of clothing segmentation. Therefore, the identity recognition and clothing segmentation problems are inter-twined; a good solution for one aides in the solution for the other.

In this chapter, we build on this idea by analyzing the mutual information between pixel locations near the face and the identity of the person to learn a global clothing mask. We segment

FIGURE 7.1: 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 clothing region in each image using graph cuts based on a clothing model learned from one or multiple images believed to be the same person wearing the same clothing. We use facial features and clothing features to recognize individuals in other images. The results show that clothing segmentation provides a significant improvement in recognition accuracy for large image collections, and useful clothing masks are simultaneously produced.

A further significant contribution is that we introduce a publicly available consumer image collection where each individual is identified. We hope this dataset allows the vision community to more easily compare results for tasks related to recognizing people in consumer image collections.

Figure 7.1 illustrates the limitations of using only facial features for recognizing people. When only six faces (cropped and scaled in the same fashion as images from the PIE [113] 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 7.1). 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 demonstrates the potential of person recognition using features in addition to the face for distinguishing individuals in family albums.

When extracting clothing features from the image, it is important to know where the clothing is located. We describe the use of graph cuts for segmenting clothing in a person image. We show that using multiple images of the same person from the same event allows a better model of the clothing to be constructed, resulting in superior clothing segmentation. We also describe the benefits of accurate clothing segmentation for recognizing people in a consumer image collection.

7.1 Related Work

Clothing for identification has received much recent research attention. When attempting to identify a person from the same day as the training data for applications such as teleconferencing and surveillance, clothing is an important cue [29, 71, 91]. In these video-based applications, good figure segmentation is achieved from the static environment.

In applications related to consumer image collections [3, 119, 124, 144, 145], clothing color features have been characterized by the correlogram of the colors in a rectangular region surrounding a detected face. For assisted tagging of all faces in the collection, combining face with body features provides a 3-5% improvement over using just body features. 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 [119] adjust the box position based on other recognized faces and attempt to exclude flesh.

Some reseachers have trained models to essentially learn the characteristics of the human form [25, 86, 104, 120]. Broadly speaking, these methods search for body parts (e.g. legs, arms, or trunk), and use a pre-defined model to find the most sensible human body amongst the detected parts. While a model-based approach is certainly justified for the problem, we wonder what can be learned from the data itself. Given many images of people, is it possible for the computer to

	Set 1	Set 2	Set 3	Set 4
Total images	401	1065	2099	227
Images with faces	180	589	962	161
No. faces	278	931	1364	436
Detected faces	152	709	969	294
Images with multiple people	77	220	282	110
Time span (days)	28	233	385	10
No. days images captured	21	50	82	9
Unique individuals	12	32	40	10

TABLE 7.1: A summary of the four image collections.

learn the shape of a human without imposing a physical human model on its interpretation of the images?

Regarding segmenting objects of interest, researchers have attemped to combine the recognition of component object parts with segmentation [140], and to recognize objects among many images by first computing multiple segmentations for each image [108]. Further, Rother *et al*. extend their GrabCut [106] graph-cutting object extraction algorithm to operate on simultaneously on pairs of images [107], and along the same lines, Liu and Chen [80] use PLSA to initialize the GrabCut, replacing the manual interface. We extend this problem into the domain of recognizing people from clothing and faces. We apply graph cuts simultaneously to a group of images of the same person to produce improved clothing segmentation.

This chapter is organized as follows: First, we analyze the information content in pixels surrounding the face to discover a global clothing mask (Section 4). Then, on each image, we use graph-cutting techniques to refine the clothing mask, where our clothing model is developed from one or multiple images believed to contain the same individual (Section 5). In contrast to some previous work, we do not use any model of the human body. We build a texture and color visual word library from features extracted in putative clothing regions of people images and use both facial and clothing features to recognize people. We show these improved clothing masks lead to better recognition (Section 7).

7.2 Images and Features for Clothing Analysis

Four consumer image collections are used in this work. Each collection owner labeled the detected faces in each image, and could add faces missed by the face detector [67]. The four

FIGURE 7.2: Person images at resolution 81×49 and the corresponding superpixel segmentations.

collections, summarized in Table 1, contain a total of 3009 person images of 94 unique individuals. We experiment on each collection separately (rather than merging the collections), to simulate working with a single person's image collection.

Features are extracted from the faces and clothing of people. Our implementation of a face detection algorithm [67] detects faces, and also estimates the eye positions. Each face is normalized in scale (61×49 pixels) and projected onto a set of Fisherfaces [11], representing each face as a 37-dimensional vector. These features are not the state-of-the-art features for recognizing faces, but are sufficient to demonstrate our approach.

For extracting features to represent the clothing region, the body of the person is resampled to 81×49 pixels, such that the distance between the eyes (from the face detector) is 8 pixels. The crop window is always axis-aligned with the image. Clothing comes in many patterns and a vast pallette of colors, so both texture and color features are extracted. A 5-dimensional feature vector of low-level features is found at each pixel location in the resized person image. This dense description of the clothing region is used based on the work of [77, 79] as it is necessary to capture the information present even in uniform color areas of clothing. The three color features are a linear transformation of RGB color values of each pixel to a luminance-chrominance space (LCC). The two texture features are the responses to a horizontal and vertical edge detector.

To provide some robustness to translation and movement of the person, the feature values are accumulated across regions in one of two ways. In the first (superpixel) representation, the person image is segmented into superpixels using normalized cuts [111], shown for example in Figure 7.2. For each superpixel, the histogram over each of the five features is computed. In

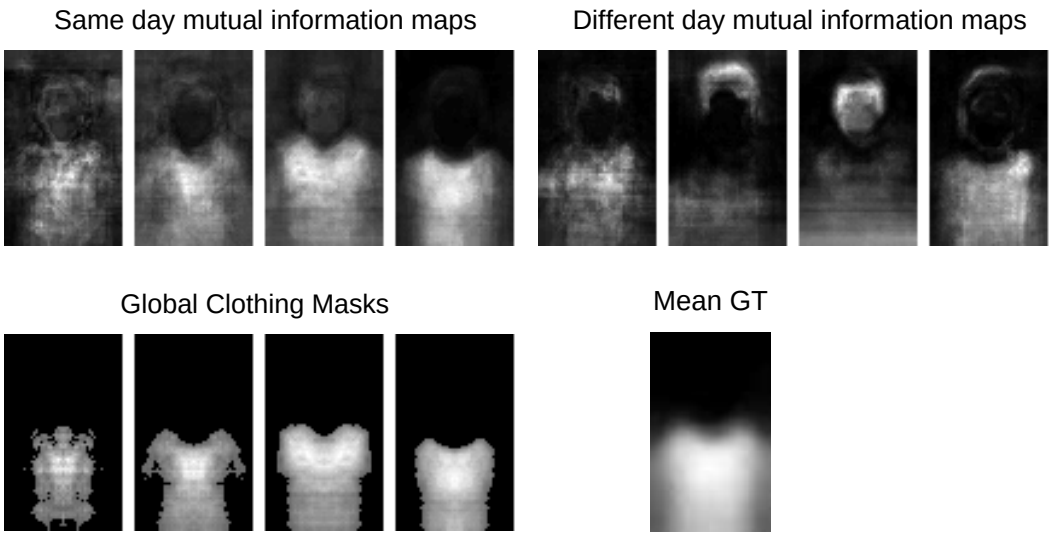


FIGURE 7.3: **Top Left:** The clothing region carries information about identity. Maps of mutual information between S_{ij} and $\langle s_{i(x,y)}, s_{j(x,y)} \rangle_s$ for four image sets all yield a map with the same qualitative appearance. In Set 3, the mutual information reaches 0.17, while the entropy of S_{ij} is only 0.19. **Top Right:** The mutual information maps for person images captured on different days. The overall magnitude is only about 7% the same-day mutual information maps, but the clothing region (and the hair region) still carry information about the identity of the person. **Bottom Left:** The clothing masks created from the mutual information masks all have the same general appearance, though Set 1's mask is noisy probably due to the relatively small number of people in this set. **Bottom Right:** The average of 714 hand-labeled clothing masks appears similar to the mutual information masks.

turn, each pixel's features are the five histograms associated with its corresponding superpixel. This representation provides localization (over each superpixel) and maintains some robustness to translation and scaling. The notation $\mathbf{s}_p$ refers to the feature histograms associated with the p^{th} superpixel. Likewise, the notation $\mathbf{s}_{(x,y)}$ refers to the feature histograms associated with the superpixel that corresponds to position (x, y).

In the second (visual word) representation, the low-level feature vector at each pixel is quantized to the index of the closest visual word [116], 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 (described in Section 4). Of course, this clothing mask is the putative region of clothing for the face; the actual clothing in a particular person image may be occluded by another object. The visual word clothing features are represented as $\mathbf{v}$.

7.3 Finding the Global Clothing Mask

In previous recognition work using clothing, either a rectangular region below the face is assumed to be clothing, or the clothing region is modeled using operator-labeled clothing from many images [117]. We take the approach of learning the clothing region automatically, using only the identity of faces (from labeled ground-truth) and no other input from the user. Intuitively, the region associated with clothing carries information about the identity of the face. For example, in a sporting event, athletes wear numbers on their uniforms so the referees can easily distinguish them. Similarly, in a consumer image collection, when two people in different images wear the same clothing, the probability increases that they might be the same individual. We discover the clothing region by finding pixel locations that carry information about facial identity. Let $p_i = p_j$ be the event S_{ij} that the pair of person images p_i and p_j share an identity, and $\langle \mathbf{s}_{i(x,y)}, \mathbf{s}_{j(x,y)} \rangle_s$ be the distance between corresponding superpixel features $\mathbf{s}_{i(x,y)}$ and $\mathbf{s}_{j(x,y)}$ at pixel position (x, y). The distance is the sum of χ^2 distances between the five feature histograms:

$$\langle \mathbf{s}_{i(x,y)}, \mathbf{s}_{j(x,y)} \rangle_s = \sum_u \chi^2(\mathbf{s}^u_{i(x,y)}, \mathbf{s}^u_{j(x,y)}) \tag{7.1}$$

where u is an index over each of the five feature types (three for color and two for texture).

In the region surrounding the face, we compute the mutual information $I(S_{ij}, \langle \mathbf{s}_{i(x,y)}, \mathbf{s}_{j(x,y)} \rangle_s)$ between the distance between corresponding superpixels, and S_{ij} at each (x, y) position in the person image. Maps of the mutual information are shown in Figure 7.3. For each image collection, two mutual information maps are found, one where p_i and p_j are captured on the same day, and one otherwise.

Areas of the image associated with clothing contain a great deal of information regarding whether two people are the same, given the images are captured on the same day. Even for images captured on different days, the clothing region carries some information about identity similarity, due to the fact that clothes are re-worn, or that a particular individual prefers a specific clothing style or color.

In three image Sets (1, 2, and 4), the features of the face region itself carry little information about identity. (Remember, these features are local histograms of color and texture features not meant for recognizing faces). These collections have little ethnic diversity so the tone of the

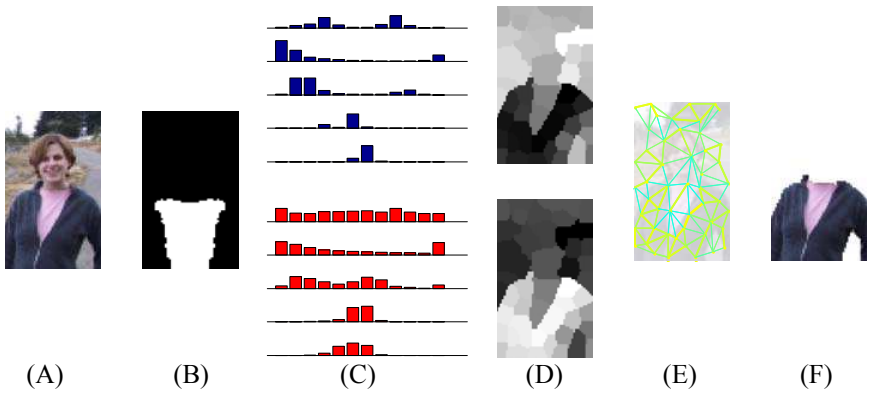


FIGURE 7.4: Using graph cuts to segment the clothing from a person image. The automatically learned global clothing mask (B) is used to create a clothing model (C, top) and a background model (C, bottom) that each describe the five feature types from the person image (A). Each superpixel is a node in a graph, and the data cost of assigning each superpixel to the clothing and background are shown (D, top) and (D, bottom), respectively, with light shades indicating high cost. The smoothness cost is shown in (E), with thicker, yellower edges indicating higher cost. The graph cut solution for the clothing is shown in (F).

facial skin is not an indicator of identity. However, Set 3 is ethnically more diverse, and the skin tone of the facial region carries some information related to identity.

This mutual information analysis allows us to create a mask of the most informative pixels associated with a face that we call the *global clothing mask*. The same-day mutual information maps are reflected (symmetry is assumed), summed, and thresholded (by a value constant across the image collections) to yield clothing masks that appear remarkably similar across collections. We emphasize again that our global clothing mask is learned without using any manually labeled clothing regions; simply examining the image data and the person labels reveals that the region corresponding roughly to the torso contains information relevant to identity.

7.4 Graph Cuts for Clothing Segmentation

Single Image: The global clothing mask shows the location of clothing on average, but on any given image, the pose of the body or occlusion can make the clothing in that image difficult to localize. We use graph cuts to extract an image-specific clothing mask. Using the idea of GrabCut [106], we define a graph over the superpixels that comprise the image, where each edge in the graph corresponds to the cost of cutting the edge. We seek the binary labeling f over the superpixels that minimizes the energy of the cut. We use the standard graph cutting algorithms [6, 18, 19, 72] for solving for the minimum energy cut. Using the notation in [72], the energy is:

$$E(f) = \sum_{p \in \mathcal{P}} D_p(f_p) + \sum_{p,q \in \mathcal{N}} V_{p,q}(f_p, f_q) \tag{7.2}$$

where $E(f)$ is the energy of a particular labeling f, p and q are indexes over the superpixels, $D_p(f_p)$ is the data cost of assigning the p^{th} superpixel to label f_p, and $V_{p,q}(f_p, fq)$ represents the smoothness cost of assigning superpixels p and q in a neighborhood $\mathcal{N}$ to respective labels f_p and f_q.

Possible labels for each superpixel are $f_p \in \{0, 1\}$ where the index 0 corresponds to foreground (i.e. the clothing region that is useful for recognition) and 1 corresponds to background. The clothing model M_0 is formed by computing the histogram over each of the five features over the region of the person image corresponding to clothing in the global clothing mask. In a similar manner, the background model M_1 is formed using the feature values of pixels from regions corresponding to the inverse of the clothing mask. Then, the data cost term in Eq. (7.2) is defined:

$$D_p(f_p) = \exp(-\alpha \langle s_p, M_{f_p} \rangle) \tag{7.3}$$

where again the distance is the sum of the χ^2 distances for each of the corresponding five feature histograms. The smoothness cost term is defined as:

$$V_{p,q}(f_p, f_q) = (f_p - f_q)^2 \exp(-\beta \langle s_p, s_q \rangle) \tag{7.4}$$

Experimentally, we found parameter values of $\alpha = 1$ and $\beta = 0.01$ work well, though the results are not particularly sensitive to the chosen parameter values. The lower value of β is explained by considering that clothing is often occluded by other image objects, and is often not contiguous in the image. Figure 7.4 illustrates the graph cutting process for segmenting the clothing region. Except for the selection of a few constants, the algorithm essentially learned to segment clothing first by finding a global clothing mask describing regions of the image with high mutual information with identity, then performing a segmentation to refine the clothing mask on any particular image.

Multiple Images: When multiple images of the same person with the same clothing are available, there is an opportunity to learn a better model for the clothing. We use the idea from [107] that the background model for each image is independent, but the foreground model is constant

across the multiple images. Then, the clothing model is computed with contribution from each of the images:

$$M_0 = \sum_i M_{0i} \tag{7.5}$$

This global clothing model M_0 is the sum for each feature type of the corresponding feature histograms for each image's individual clothing model. However, each image i has its own individual background model M_{1i}, formed from the feature values of the inverse global clothing mask. Conceptually, the clothing is expected to remain the same across many images, but the background can change drastically.

When applying graph cuts, a graph is created for each person image. The smoothness cost is defined as before in Eq. (7.4), but the data cost for person image i becomes:

$$D_{pi}(f_{pi}) = \begin{cases} \exp(-\alpha\langle s_{pi}, M_0\rangle) & \text{if } f_{pi}\text{=0} \\ \exp(-\alpha\langle s_{pi}, M_{1i}\rangle) & \text{if } f_{pi}\text{=1} \end{cases} \tag{7.6}$$

Figure 7.5 shows several examples of graph cuts for clothing segmentation by either treating each image independently, or exploiting the consistency of the clothing appearance across multiple images for segmenting each image in the group.

7.5 Recognizing people

For searching and browsing images in a consumer image collection, we describe the following scenario. At first, none of the people in the image collection are labeled, though we do make the simplifying assumption that the number of individuals is known. A user provides the labels for a randomly selected subset of the people images in the collection. The task is to recognize all the remaining people, and the performance measure is the number of correctly recognized people. This measure corresponds to the usefulness of the algorithm in allowing a user to search and browse the image collection after investing the time to label a portion of the people. 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 $\mathbf{f} = \{\mathbf{f}^f, \mathbf{v}\}$ includes the facial features $\mathbf{f}^f$ and the clothing features $\mathbf{v}$, the probability that the name assigned to person p is n is estimated using

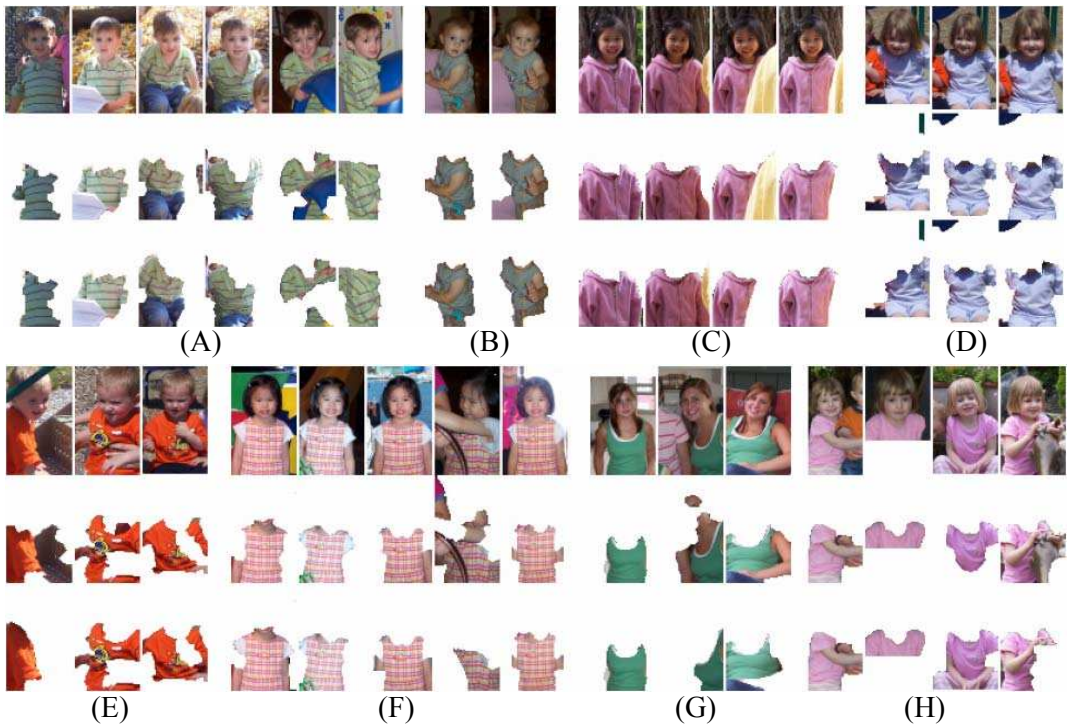


FIGURE 7.5: See Section 5. For each group of person images, the top row shows the resized person images, the middle row shows the result of applying graph cuts to segment clothing on each person image individually, and the bottom row shows the result of segmenting the clothing using the entire group of images. Often times, the group graph cut learns a better model for the clothing, and is able to segment out occlusions (A, C, F, H) and adapt to difficult poses (E, G). We do not explicitly exclude flesh, so some flesh remains in the clothing masks (B, G, H).

nearest neighbors. In our notation, name set $\mathbf{N}$ comprises the names of the U unique individuals in the image collection. An element $n^k \in \mathbf{N}$ is a particular name in the set. The K nearest labeled neighbors of a person p_i are selected from the collection using facial similarity and clothing similarity. When finding the nearest neighbors to a query person with features $\mathbf{f}$, 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:

$$P_{ij} = P(S_{ij}|\mathbf{f}_i, \mathbf{f}_j, t_i, t_j) \tag{7.7}$$

$$\approx \max\left[P_{ij}^v, P_{ij}^f\right] \tag{7.8}$$

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, similar to the superpixel distance in Eq. (7.1). 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. Figure 7.9 shows the maximum likelihood estimate of P_{ij}^v. The posterior is fit with a decaying exponential, one model for person images captured on the same day, and one model for person images captured on different days. Similarly, the probability P_{ij}^f, the probability that faces i and j are the same person, is modeled using a decaying exponential.

We justify the similarity metric P_{ij} based on our observations of how humans perform recognition by combining multi-modal features to judge the similarity between faces. 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 judgement 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). An example of this phenomena is shown in Figure 7.10.

Using the metric P_{ij}, a nearest neighbor is one that is similar in either facial appearance or in clothing appearance. These K nearest neighbors are used to estimate $P(p = n|\mathbf{f})$ using a weighted density estimate, which can in turn be used to recognize the face according to:

$$p_{\text{MAP}} = \arg\max_{n \in \mathbf{N}} P(p = n|\mathbf{f}) \tag{7.9}$$

When multiple people are in an image, there is an additional constraint, called the *unique object constraint*, that no person can appear more than once in an image [13, 117]. We seek the assignment of names to people that maximizes $P(\mathbf{p} = \mathbf{n}|\mathbf{F})$, the posterior of the names for all people in the image, assuming that any group of persons is equally likely. The set of M people in the image is denoted $\mathbf{p}$, $\mathbf{F}$ is the set of all the features $\mathbf{f}$ for all people in the image, and $\mathbf{n}$ is a subset of $\mathbf{N}$ with M elements and is a particular assignment of a name to each person in $\mathbf{p}$. Although there are $\binom{U}{M}$ combinations of names to people, this problem is solved in $O(M^3)$ time using Munkres algorithm [87].

FIGURE 7.6: A demonstration of clothing retrieval. For each row, the image on the left is the query image. The remaining seven images are the closest people, based on the clothing distance, in the image collection.

FIGURE 7.7: Combining color and texture for clothing representation improves the results. For each row, the query image is on the left. **Top Row:** Clothing retrieval using color features only. **Middle Row:** Clothing retrieval using texture only. **Bottom Row:** Clothing retrieval with both color and texture provides the best result.

7.6 Retrieval

The clothing distance $d(\mathbf{f}_i^c, \mathbf{f}_j^c)$ can be used for the task of clothing retrieval, where the goal is to sort all of the person images based on the similarity of the clothing region associated with a detected face. Figure 7.6 shows several examples of clothing retrieval, which is a useful

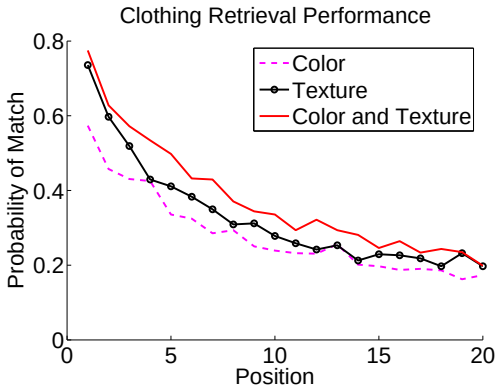


FIGURE 7.8: The performance of clothing retrieval on a set of 715 labeled people with 50 different clothing items. Nearly 80% of the time the best match has the same clothing as the query.

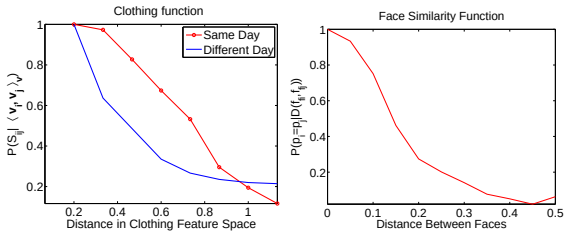


FIGURE 7.9: **Left:** The probability that two person images share a common identity given the distance between the clothing features and the time interval between the images. **Right:** In a similar fashion, the probability of two person images sharing a common identity given the distance between faces $\mathbf{f}_i^f$ and $\mathbf{f}_j^f$.

application in its own right. For example, a fashion editor may use such a system to discover two celebrities wearing the same outfit to an event. In Figure 7.7 the advantage of using both color and texture visual words for retrieval is demonstrated with an example. As an experiment, each of the 715 labeled clothing items in turn is considered as the query, and the closest 15 people images from the entire image collection are returned. The retrieval performance, as a function of position in the search results, is shown in Figure 7.8. As expected, using both the color and texture visual words for retrieval is better than using either color or texture visual words alone. Recall that this clothing retrieval is accomplished based only on the automatic detection of face position; the user never circumscribes or otherwise indicates which areas of the query image contain clothing.

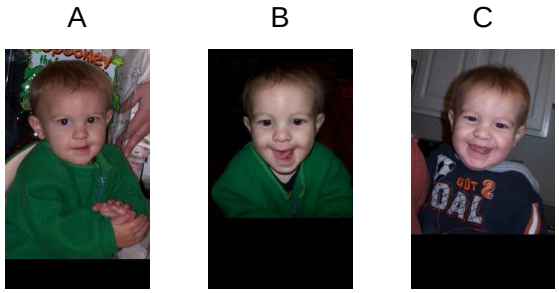


FIGURE 7.10: Three images of the same baby as a justification for the measure of similarity P_{ij}. Persons A and B (captured on the same day) have high clothing similarity, yet the differences in facial pose and expression result in low facial similarity. We trust the clothing similarity for this pair. Likewise, persons B and C have low clothing similarity, but high facial similarity (both faces have the same pose and similar open-mouth smile. The low clothing score does not affect our belief that person B and C are the same, based on the facial similarity. A and C are dissimilar in both facial measurements and clothing. Yet, by transitivity, we conclude A and C are likely to be the same individual.

7.7 Discovering Clusters of People

The pairwise posterior probabilities of people having the same identities given clothing features P_{ij}^c or facial features $P_{ij}^f = P(S_{ij}|d(\mathbf{f}_i^f, \mathbf{f}_j^f))$ can be thought of as edge weights on a complete graph, where each person is a vertex. People images in an image collection can be automatically grouped to discover clusters using graph segmentation algorithms such as normalized cut [111], as for example has been demonstrated to find actors in movies [46] using facial similarity.

When the edge weights are formed by P_{ij}^{cf}, clusters of people emerge containing people with similarities in face and clothing, but not necessarily both simultaneously. Normalized cut seeks to maximize the association within clusters, and minimize the cut between clusters. Intuitively, each person in a cluster is not necessarily similar to all the other persons in the cluster, but has similarity (in face or clothing) with one or more other persons in the cluster. In practice, the clusters often have a semantic meaning that is easy to recognize, provided an appropriate number of clusters are used. Figure 7.11-7.13 shows some example clusters that emerge.For this example, we use 20 clusters for faces, 40 for clothing, and 80 for the combination of face and clothing.

7.8 Experiments

Better Recognition Improves Clothing Segmentation: The following experiment was performed to evaluate the performance of the graph-cut clothing segmentation. In our Sets 1 and 4,

FIGURE 7.11: Three clusters of people with facial similarity.

every superpixel of every person image was manually labeled as either clothing or not clothing. This task was difficult, not just due to the sheer number of superpixels (35700 superpixels), but because of the inherent ambiguity of the problem. For our person images, we labeled as clothing any covering of the torso and legs. Uncovered arms were not considered to be clothing, and head coverings such as hats and glasses were also excluded.

We apply our clothing segmentation to each person image in both collections. Table 2 reports the accuracy of the clothing segmentation. We compare the graph cut segmentation against the prior (roughly 70% of the superpixels are *not* clothing). A naïve segmentation is to find the mean value of the clothing mask corresponding to the region covered by each superpixel, then classify as clothing if this value surpasses a threshold. The threshold was selected by minimizing the equal error rate. This method considers only the position of each superpixel and not its feature values. In both collections, using the graph cut clothing segmentation provides a substantial improvement over the naïve approach.

FIGURE 7.12: Five clusters of people with clothing similarity.

	Set 1	Set 4
Prior	70.7%	68.2%
Naïve	77.2%	84.2%
GC Individual	87.6%	88.5%
GC Group	**88.5**%	**90.3**%

TABLE 7.2: Graph cuts provides effective clothing recognition. For each of two image collections, the accuracy of classifying superpixels as either clothing or non-clothing with four different algorithms is shown. Using Graph Cuts for groups of images proves to be the most effective method.

Further improvement is achieved when the person images are considered in groups. For this experiment, we assume the ground truth for identity is known, and a group includes all instances of an individual appearance within a 20 minutes time window, nearly ensuring the clothing has not been changed for each individual.

Better Clothing Recognition Improves Recognition: The following experiment is performed to simulate the effect on recognition of labeling faces in an image collection. People images are labeled according to a random order and the identity of all remaining unlabeled faces is inferred by the nearest-neighbor classifier from Section 6. Each classification is compared against the true label to determine the recognition accuracy. We use nine nearest neighbors and repeat

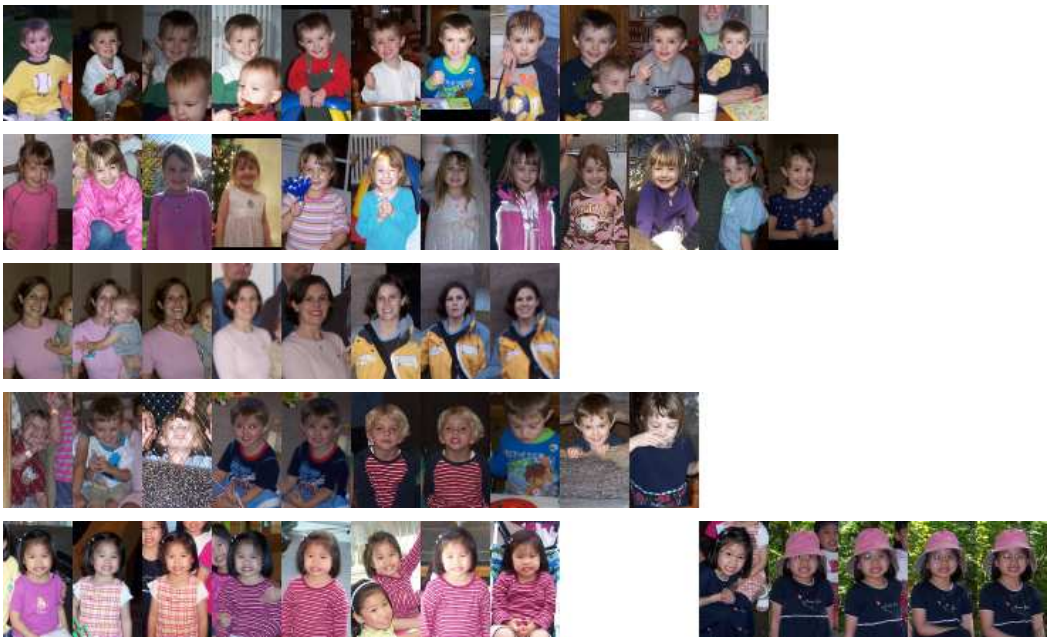


FIGURE 7.13: Six clusters of people using an affinity matrix constructed with entries P_{ij}^{cf}.

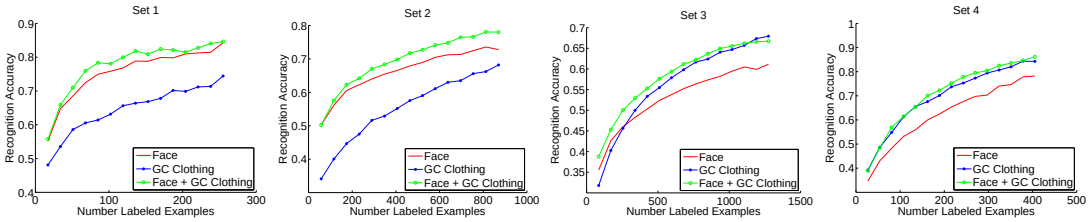


FIGURE 7.14: Combining facial and clothing features results in better recognition accuracy than using either feature independently.

the random labeling procedure 50 times to find the average performance. The goal of these experiments is to show the influence of clothing segmentation on recognition.

Figure 7.14 shows the results of the person recognition experiments. The combination of face and clothing features improves recognition in all of our test sets. If only a single feature type is to be used, the preferred feature depends on the image collection. For this experiment, the clothing features are extracted from the clothing mask determined by graph cuts on each image individually.

Figure 7.15 compares the performance of recognizing people using only clothing features. For all of our collections, the graph cut clothing masks outperform using only a box (shown in Figure 7.16). Also, for each collection, the clothing masks are generated by segmenting using

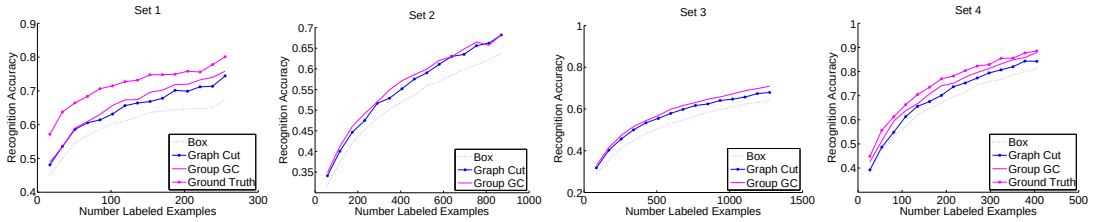


FIGURE 7.15: Using graph cuts for the extraction of clothing features improves the accuracy of recognizing people over using a simple box region. Further improvement is attained by using multiple person images when performing clothing segmentation. Sets 1 and 4 demonstrate even more room for improvement when ground-truth clothing segmention is used for feature extraction.

group segmentation, and these segmentations unanimously lead to better recognition performance. Finally, we show in collection Sets 1 and 4, where ground-truth labeled clothing masks exist, that the best performance is achieved using the ground truth clothing masks. This represents the maximum possible recognition accuracy that our system could achieve if the clothing segmentation is perfect.

To summarize, these experiments show that:

- Multiple images of the same person improve clothing segmentation.
- Person recognition improves with improvements to the clothing segmentation.

Ongoing work includes merging the recognition and clothing segmentation into a single framework where each assists the other in the following fashion: based on a labeled subset of people, the other people in the collection are recognized. Then, based on these putative identities, new clothing masks are found using multiple images of the same person within a given time window.

7.9 Publically Available Dataset

One persistant problem for researchers dealing with personal image collections is that there is a lack of standard datasets. As a result, each research group uses their own datasets, and results are difficult to compare. We have made our image Set 2 of 931 labeled people available to the research community [47]. The dataset is described in Table 1, and contains original JPEG captures with all associated EXIF information, as well as text files containing the identity of all labeled individuals. We hope this dataset provides a valuable common ground for the research community.

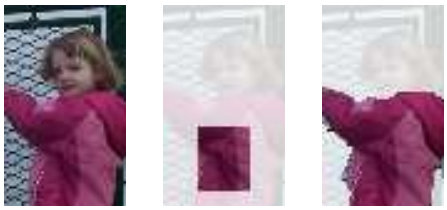


FIGURE 7.16: Given an image (left), using the clothing features from a graph cut clothing mask (right) results in superior recognition to using a box (middle).

7.10 Conclusion

In this chapter, we describe the advantages of performing clothing segmentation with graph cuts in a consumer image collection. We showed a data-driven (rather than driven by a human model) approach for finding a global clothing mask that shows the typical location of clothing in person images. Using this global clothing mask, a clothing mask for each person image is found using graph cuts. Further clothing segmentation improvement is attained using multiple images of the same person which allows us to construct a better clothing model.

This work can be viewed as a case study for the merits of combining segmentation and recognition. Improvements in clothing segmentation improve person recognition in consumer image collections. Likewise, using multiple images of the same person improves the results of clothing segmentation.

Chapter 8

Using Group Prior to Identify People in Consumer Images

While face recognition techniques have rapidly advanced in the last few years, most of the work is in the domain of security applications. For consumer imaging applications, person recognition is an important tool that is useful for searching and retrieving images from a personal image collection. It has been shown that when recognizing a single person in an image, a maximum likelihood classifier requires the prior probability for each candidate individual. In this paper, we extend this idea and describe the benefits of using a group prior for identifying people in consumer images with multiple people. The group prior describes the probability of a group of individuals appearing together in an image.

In our application, we have a subset of ambiguously labeled images for a consumer image collection, where we seek to identify all of the people in the collection. We describe a simple algorithm for resolving the ambiguous labels. We show that despite errors in resolving ambiguous labels, useful classifiers can be trained with the resolved labels. Recognition performance is further improved with a group prior learned from the ambiguous labels. In summary, by modeling the relationships between the people with the group prior, we improve classification performance.

Figure 8.1 shows a few example images containing people from a single image collection. Because each person is a unique individual, we immediately have a powerful constraint that affects the design and selection of a classifier. Within an image, an individual can appear at most one time, and each person in an image can be only one individual [13]. (We ignore the rare images

FIGURE 8.1: Example of a few images from an image collection. Ambiguous labels provide the information about who is in each image and are used to estimate the group prior.

containing a face and its mirror reflection, or images containing other images, etc.) This intuitive constraint provides a foundation for determining the identities of people from consumer images. We call this constraint the *unique object constraint*.

When multiple people are in an image, there is usually a relationship between the people in the image. For example, the people could be friends, co-workers, siblings, or relatives. By learning the prior probability of different individuals appearing together in an image, classification can be improved. This prior probability of certain groups of people appearing in an image is called the *group prior*. The group prior implicitly incorporates the unique object constraint, because the probability of any person appearing more than once in an image is zero.

Ambiguous labels are sometimes supplied with a set of images containing people. An ambiguous label provides a label for a unique object that appears in an image, without indicating which object is associated with which label. Figure 8.1 shows the ambiguous labels associated with several images. Ambiguous labels for individuals' names in images occur naturally in several situations. First, many software packages (e.g. www.flickr.com) allow the user to tag images with any keyword related to the image. Second, many people annotate their images with captions such as "George and Martha in their canoe" which conveys that Martha and George are in the image but does not indicate which is George and which is Martha. We seek to resolve the ambiguous labels by assigning each label to a specific face in the image. In addition, ambiguous labels provide exactly the information we need to estimate the group prior, which can be used to improve classification performance.

In this paper, we present algorithms that incorporate the group prior to model the relationships between people in the images. In Section 2, we review the related work. We describe a database

for recognizing people in consumer images in Section 3. We then describe an algorithm to resolve ambiguous labels (Section 4). Finally, in Section 5, we show how labeling a small image set with ambiguous labels can be used to learn group prior information and train classifiers that recognize faces in previously unseen and unlabeled images for the purpose of automatic annotation or retrieval.

8.1 Related Work

Certainly, there are many techniques for recognizing faces, or for comparing the similarity of two faces [147]. However, there are many significant differences between the problem of face recognition in general and the problem of recognizing people in consumer images. The field of face recognition emphasizes the development of features that are useful for recognition, and generally ignores issues related to prior probabilities (of an individual or specific group of individuals appearing in an image.)

With regard to capitalizing on problem-specific constraints, several classification and clustering algorithms have been developed that either implicitly or explicitly examine constraints to improve the performance of the classifier. In unsupervised clustering, Wagstaff *et al*. [130], describe an algorithm that uses known constraints between example points. The "must-link" constraint requires that two examples be in the same cluster while the "cannot-link" constraint requires that the two points cannot be in the same cluster. Constraints have also been added to clustering algorithms such as normalized cut [94, 111]. The constraints can relate to lane segmentation [130], image segmentation [141], or inferring web page relationships [111]. When considering faces from many images, all faces from a single image are all mutually "cannot-link" due to the unique object constraint and there are no "must-link" constraints. These approaches do consider the problem constraints, but they do not incorporate labeled data and are not suitable for our application.

Computer vision researchers have worked with ambiguously labeled data. Satoh and Kanade [109] developed the "Name-It" system to identify faces in news video from the transcripts and video captions. Berg *et al*. [13] extract names from captions of news photos and associate the names with faces in the images. Both these applications involve noisy labels (i.e. a detected name may not be someone who appears in the image) and are difficult problems. Berg handles this noise by initializing the name-face assigment algorithm using those images containing only

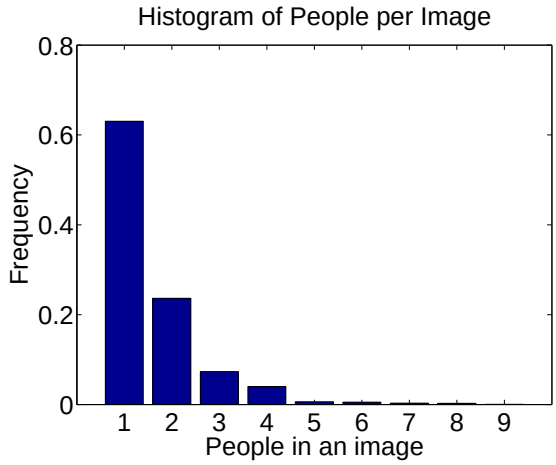


FIGURE 8.2: **Left:** A histogram of the number of people per image from a set of four image collections of over 3500 consumer images.

a single face with only one name in the caption, then uses expectation maximization to assign names to faces. Our ambiguously labeled images are related to this work, but we assume that a human is actively providing the ambiguous labels for each image's detected faces. Thus, we expect that a name provided by the human will appear in the image, and therefore avoid the noisy label problem.

In an example of using weakly labeled data, Zhang *et al*. [145] describe a photo organizing system where a user indicates a set of images that contain a certain person, and the system selects one face from each of the images that maximize the overall similarity between the selected faces.

Our work builds on these techniques by improving the recognition performance using a group prior. The group prior serves as the context for the classification problem, akin to performing object detection by setting the context of the scene [126]. The cooccurance of individuals in images has been considered by Naaman *et al*. [90] for an interactive image labeling application that uses only image context (like the image capture time and place, and other people in the image) to suggest the next most likely label name for an image. We build on the work of Naaman *et al*. by finding the prior for any group or people (rather than single person) in the image, and combining that prior with facial features. Our work extends that of Zhang *et al*. by simultaneously handling multiple person names to disambiguate the ambiguous labels. Our ambiguous label resolution algorithm handles a simpler problem than either [13, 109] yet it does not need to be initialized with faces having known labels. In summary, classification is improved by considering the features of all people in the image along with the group prior.

8.2 Images and Features

Much of the work described in this paper takes advantage of constraints that naturally occur when multiple persons appear in a single image. Therefore, it is important to understand the distribution of people in images.

Four image collections were acquired, containing a total of 1084 images with people. Each collection owner labeled the people in each image. The database includes 1924 labeled instances of 114 unique people. Analysis of the collected face identities provides a rich set of information for recognition algorithm development. Figure 8.2 shows a histogram of the number of people in an image in images with people. About 50% of the images containes one or more people, and of these many contain more than one person. Each image collection has a small number of people that appear very often. These popular people are the ones we would like to be able to recognize, as they are obviously important to the photographer. In our image collections, the number of popular people ranges from five to eleven.

A face detection algorithm [67] is used to detect faces in each image. Facial features based on facial geometry are robust to some variation in pose and illumination that is typically encountered in consumer photography [147]. An active shape model [30] is used to locate the positions of 82 key points for each face, and each face is represented as a 5-dimensional feature vector. An example face having the automatically determined key points is shown in Figure 8.3. These features are not the state-of-the-art features for recognizing faces, but are sufficient to demonstrate our approach.

The feature vectors associated with faces from an image collection can be visualized by plotting each face according to the first two dimensions of the feature space, as shown in Figure 8.4. Each individual's feature vectors are plotted with a different symbol. We are interested in studying the group prior with images containing more than one of the popular unique individuals. The four image collections contain 61, 204, 420, and 455 faces with at least two faces per image, and 5, 5, 5 and 11 popular unique individuals respectively. In Figure 8.4, a line is drawn between faces that appear in the same image. This corresponds with the unique object constraint that since an individual can only appear once in an image, any two faces joined by a line must be different individuals. The image collections have 44, 237, 288, and 360 total constraints, respectively. Each constraint is related to a unique pair of faces in an image. Table 1 summarizes these datasets which are used throughout this paper.

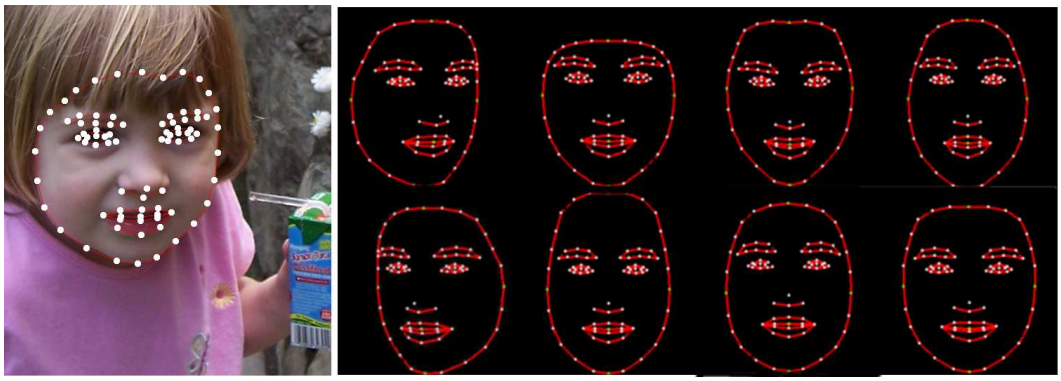


FIGURE 8.3: **Left:** An image with 82 key points automatically identified. **Right:** PCA is used to represent each face with a 5-dimensional feature vector, corresponding to eigenvectors that relate to differences in individual appearance. The visualization of the first four eigenvectors of the key points is shown. The top row corresponds to the average face plus the eigenvector, and in the bottom row the eigenvector is subtracted from the average face. The first and third eigenvectors relate to facial pose and are ignored. The second and fourth eigenvectors relate to differences in individual appearance and are preserved.

	Set 1	Set 2	Set 3	Set 4
Total images	300	300	1197	2099
Images with multiple people	26	67	188	191
No. faces from these images	61	204	420	455
Constraints	44	237	288	360
Popular unique individuals	5	5	5	11

TABLE 8.1: Information about the four datasets.

8.3 Resolving Ambiguous Labels

An ambiguously labeled image has associated individual names but the labels do not indicate which person is which individual. The caption of Figure 8.1 gives an example of ambiguous labels for three images. Once the ambiguous labels have been resolved, we have a collection of labeled faces. A classifier can be trained with these labels so that faces from completely unlabeled images from the same collection can be recognized. Figure 8.5 illustrates the proposed system.

We resolve the ambiguous labels by assigning each label to a person in an image. The objective function is the sum of squared distances between each face and the associated cluster center for its label. Certainly, minimizing this objective function by computing it for every possible assignment of labels is out of the question for all but the smallest number of faces and images.

Given a set of J ambiguously labeled images, the goal is to assign each face to a cluster C_k corresponding to one of K label names in the name set $\mathbf{N}$ (where K is the number of unique

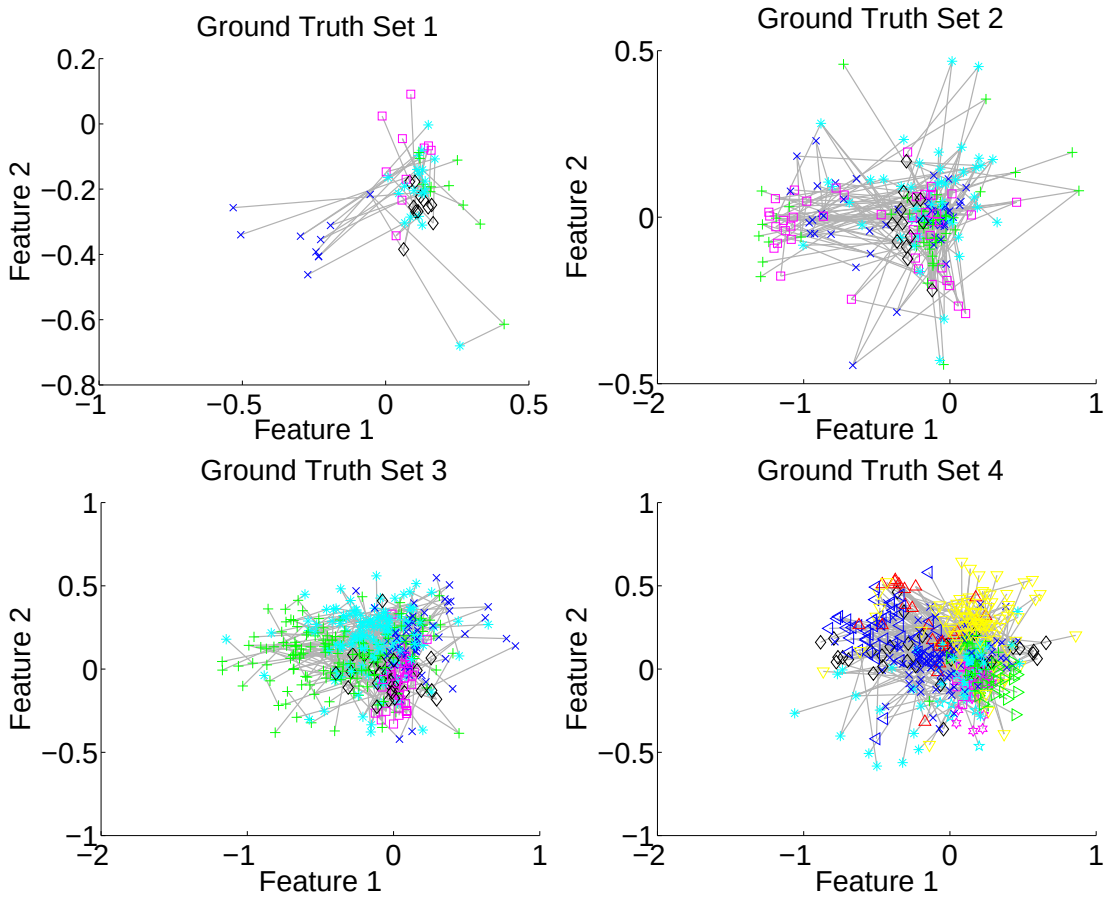


FIGURE 8.4: The four test image collections. Each data point represents a face (projected to the first two feature dimensions). Each unique symbol represents a different individual in that image collections. Lines connect faces that appear in the same image.

names among the ambiguous labels.) Let f_{mj} represent the features for the m^{th} face from the j^{th} image. M_j is the number of faces in the j^{th} image. Every image with more than one face has a unique object constraint that f_{mj} and f_{nj} cannot belong to the same cluster C_k, $\forall m \neq n$. An element $n^k \in \mathbf{N}$ is a particular name in the set. The notation n^k_m indicates that the name n^k is associated with person m from an image. In addition to the unique object constraint, we have an additional constraint that each image's faces can only be assigned to a subset of the possible labels $\mathbf{N}$ (the ambiguous labels for that image). For image j, the ambiguous labels are $\Psi_j \subseteq \mathbf{N}$.

An algorithm for resolving ambiguous labels is ALR:

ALR: Ambiguous Label Resolution Algorithm

1. For each image j, randomly assign faces f_{mj} to ambiguous labels Ψ_j.

2. Compute the parameters of each label's cluster from the faces assigned to that label.

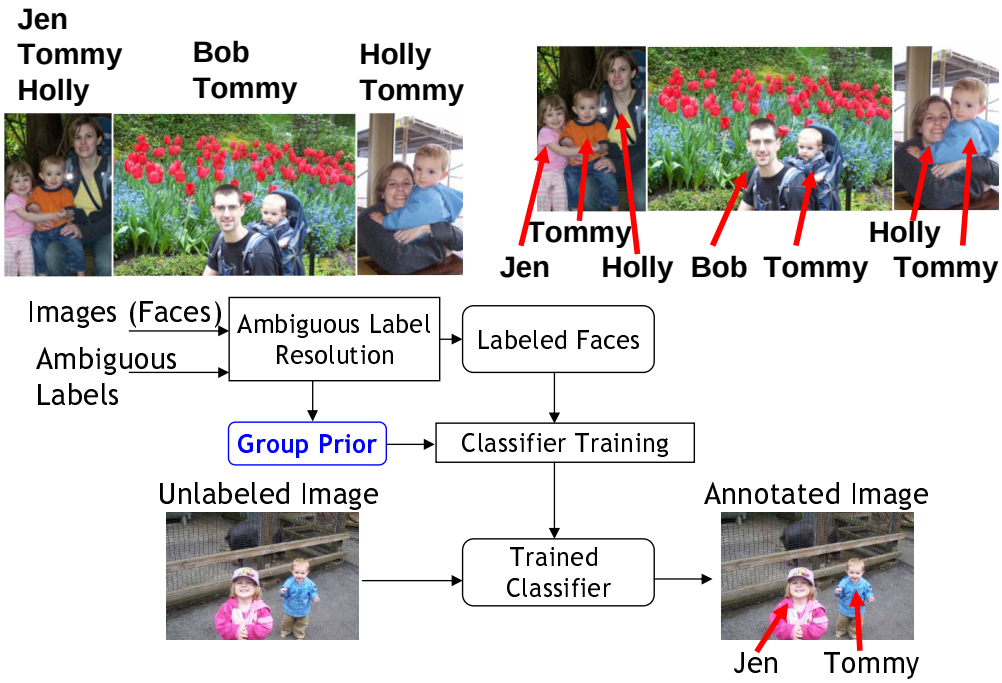


FIGURE 8.5: A system diagram. An image collection with ambiguous labels first has the ambiguous labels resolved. Then, a classifier is trained for each individual and the group prior is learned. Finally, faces in unlabeled images are classified.

3. For each image j, assign faces f_{mj} to labels Ψ_j in a manner that respects the unique object constraints and minimizes the overall squared distance E_j for the image, using the Hungarian algorithm [87].

4. Iterate between 2 and 3 until convergence.

5. Return the final assignments of faces to clusters.

Step 3 requires further explanation. For each image j, we assign all faces from that image to ambiguous labels Ψ_j such that the sum of squared distances between each face and the corresponding cluster center C_k is minimized. We construct the matrix D, having elements d_{mk} where d_{mk} is the squared distance from the m^{th} face to the k^{th} cluster center, and $k \in \Psi_j$. Then, the Hungarian algorithm is used to find the optimal assignment of faces to clusters (in polynomial time) that minimizes the overall squared distance E_j for the image. The residual error for the j^{th} image is $E_j = \sum_{m,k}^{M_j} z_{mk} d_{mk}$, where z_{mk} is an indicator variable that is 1 when face m is assigned to cluster k and 0 otherwise. As an alternative to representing each cluster by its centroid, each cluster can be described as a Gaussian, but for our data, the resolved labels were not significantly different. The key is not necessarilary how we represent each distribution, but how each face is assigned to a cluster.

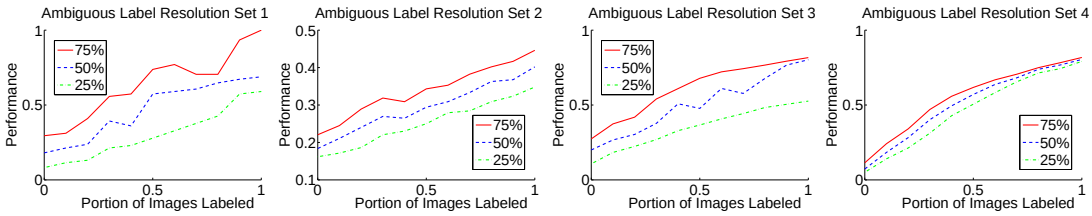


FIGURE 8.6: Performance of Ambiguous Label Resolution. The graphs show the median, 25% and 75% performances from 150 trials on each of four image collections as a function of the portion of the image collection that was ambiguously labeled.

8.3.1 Evaluation

The ambigous label resolution algorithm was applied to four consumer image collections. A portion of the images are randomly selected to be ambiguously labeled. Figure 8.12 shows an example of the resolved ambiguous labels for a set of images. The performance of the algorithm is quantified by finding the fraction of the number of all faces that are assigned the correct labels, and the results of a set of 150 trials with random initialization are shown in Figure 8.6. As expected, the performance of the algorithm improves as the number of ambiguously labeled images increases. It should be noted that the ALR algorithm, like k-means, is sensitive to the initial starting condition. In practice, multiple restarts are used and the start which converges to the minimum objective function is returned [89]. With our data ALR always converged, generally in fewer than 20 iterations.

8.4 Classifying with Resolved Labels

Using the resolved labels, a classifier is trained for recognizing the individuals in the image collection. Of course, the resolved ambiguous labels contain some errors, so we must determine whether an effective classifier can be designed in the face of these erroneously labeled samples. We make the assumption that the people in the unlabeled images are in the set $\mathbf{N}$ of the unique individuals.

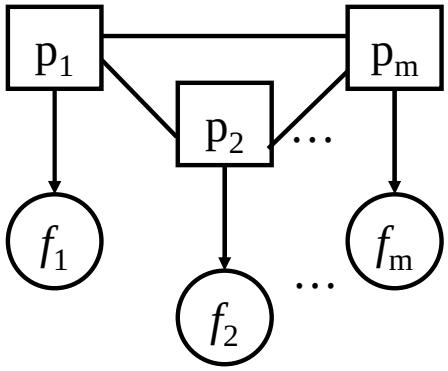


FIGURE 8.7: A graphical model that represents the features f and the people p in an image. Each person p_m has an undirected connection to all other people.

8.4.1 Images with one face

When an unlabeled image contains only a single face, the label for the face with features f is found according to Bayes rule:

$$p_{\text{MAP}} = \arg\max_{n \in \mathbf{N}} P(n|f) \tag{8.1}$$

$$= \arg\max_{n \in \mathbf{N}} P(f|n)P(n) \tag{8.2}$$

The distribution $P(f|n)$ is modeled with a Gaussian. When the computed covariance matrix is ill-conditioned, a generic covariance matrix, derived from many individuals, is substituted as the covariance matrix for that label. We have only ambiguously labeled images, so the Gaussians are computed using the resolved ambiguous labels.

The estimate of the prior probability $P(n)$ is derived from the ambiguous labels by counting the number of images containing a specific individual, according to:

$$P(n) = \frac{\sum_j y_{nj}}{\sum_u \sum_j y_{uj}} \tag{8.3}$$

where

$$y_{nj} = \begin{cases} 1 & n \in \Psi_j \\ 0 & \text{otherwise} \end{cases} \tag{8.4}$$

8.4.2 Images with multiple faces

The identities of multiple people in an image are not independent. There are two intuitive reasons for this. First, according to the unique object constraint, each individual can only appear once in the image. Second, multiple people in a consumer image generally have some kind of personal relationship. The group prior represents both the unique object constraint and the relationship between individuals that makes one group more likely to appear together in an image than another. For example, if we believe that Jen is in the image, then our belief that her brother Tommy is also in the image might increase. Thus, the classification of face identity should consider the features associated with all faces in the image.

Figure 8.7 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 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:

$$P(\mathbf{p} = \mathbf{n}|\mathbf{f}) = \frac{P(\mathbf{f}|\mathbf{p} = \mathbf{n})P(\mathbf{p} = \mathbf{n})}{P(\mathbf{f})} \tag{8.5}$$

$$\propto P(\mathbf{p} = \mathbf{n}) \prod_m P(f_m|p_m = n_m) \tag{8.6}$$

Consistent with the model, we proceed from (8.5) to (8.6) 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. Tommy looks like Tommy regardless of who else is in the image.

Because we have access to a set of ambiguously labeled images, we can estimate the group prior $P(\mathbf{p} = \mathbf{n})$ or equivalently $P(\mathbf{n})$, the prior probability that a particular set $\mathbf{n}$ of M individuals would appear together in an image. First, we consider the case of estimating the group prior for any combination of two individuals:

$$P(n^u, n^v) = \frac{\sum_j y_{n^u j}\, y_{n^v j} + \alpha(u, v)}{\sum_{g,h \in N} \sum_j y_{n^g j}\, y_{n^h j} + \alpha(g, h)} \tag{8.7}$$

where

$$\alpha(u,v) = \begin{cases} \beta & u \neq v \\ 0 & \text{otherwise} \end{cases} \tag{8.8}$$

The function $\alpha(u,v)$ with a small non-zero β ensures that any two people have a non-zero probability of appearing together and at the same time respects the unique object constraint. The prior is estimated by counting the number of images that the pair n_u and n_v appear in together, divided by the total number of pairs of people in all images. One beautiful aspect is that this estimate is independent of the outcome of the ambiguous label resolution algorithm, so $P(n^u, n^v)$ is the maximum likelihood estimate of the group prior. The size of $P(\mathbf{n})$ grows exponentially with the number of elements in $\mathbf{n}$, yet Figure 8.2 shows that images with increasing numbers of people are more rare. Instead of attempting to learn $P(\mathbf{n})$ for large M (i.e. $M > 2$) from the data, we estimate it from $P(n^u, n^v)$:

$$P(\mathbf{n}) = \frac{\prod_{u,v \in \mathbf{n}} P(n^u, n^v)}{\sum_{\mathbf{q} \subseteq \mathbf{N}} \prod_{u,v \in \mathbf{q}} P(n^u, n^v)} \tag{8.9}$$

where $\mathbf{q}$ has M elements. Equation (8.9) 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 the model of Figure 8.7.

For a particular image with M people in an image collection of K unique individuals, there can be $\text{Vals}(\mathbf{n})$ different assignments of names to the people in the image.

$$\text{Vals}(\mathbf{n}) = \binom{K}{M} M! \tag{8.10}$$

$\text{Vals}(\mathbf{N})$ grows exponentially with both K and M, so we are relieved that both tend to be small so we can explicitly solve for $P(\mathbf{p} = \mathbf{n}|\mathbf{f})$. For example, when $K = 7$ and $M = 5$, $\text{Vals}(\mathbf{N}) = 2520$.

Once $P(\mathbf{p} = \mathbf{n}|\mathbf{f})$ is found, there are many different inference questions that can be answered by marginalizing the joint distribution.

8.4.2.1 Most Probable Explanation (MPE)

In MPE, the goal is to find the most probable labeling of all faces in the image. This assignment corresponds to the mode of $P(\mathbf{p} = \mathbf{n}|\mathbf{f})$:

$$\mathbf{p}_{\text{MPE}} = \arg\max_{\mathbf{n}\in\mathbf{N}} P(\mathbf{p} = \mathbf{n}|\mathbf{p}) \tag{8.11}$$

8.4.2.2 Maximum Apriori Probability (MAP)

In MAP, the goal is to find the most probable identity of the m^{th} particular individual in the image. Therefore we marginalize over the name assignments of the other $M - 1$ people in the image.

$$p_{m\text{MAP}} = \arg\max_{n_m^k\in\mathbf{N}} \sum_{p_i, i\neq m} P(\mathbf{p} = \mathbf{n}|\mathbf{f}) \tag{8.12}$$

8.4.2.3 Ambiguously Labeling

Inference can provide ambiguous labels for an unlabeled image. We desire to name the individuals in the image, but we do not specify which face is associated with which name. This would be particularly useful for auto-captioning the image.

$$p_{\text{AMB}} = \arg\max_{\mathbf{n}\in\mathbf{N}} \sum_{\mathcal{P}(\mathbf{n})} P(\mathbf{p} = \mathbf{n}|\mathbf{f}) \tag{8.13}$$

where $\mathcal{P}(\mathbf{n})$ denotes all permutations of the set $\mathbf{n}$.

8.4.2.4 Retrieval Based on Identity

Perhaps the most important query that could be posed is: Given the observed features f, what is the probability that a particular person n^q is in this image? This query has obvious applications for image retrieval based on whom the image contains. A query for images of a particular person can return images ranked according to $P(n^q|\mathbf{f})$. To satisfy this query, we simply sum $P(\mathbf{p} = \mathbf{n}|\mathbf{f})$ over all sets of $\mathbf{n}$ where one p_m is assigned to n_m^q.

$$P(n^q|\mathbf{f}) = \sum_{\mathbf{n}, n^q\subset\mathbf{n}} P(\mathbf{p} = \mathbf{n}|\mathbf{f}) \tag{8.14}$$

8.4.3 Evaluation

Classification with the group prior was applied to the four image collections. Facial geometry features were extracted as described. One image is selected as the test image. A portion of the remaining images are ambiguously labeled and input to ALR. The group prior $P(\mathbf{n})$ is estimated from the ambiguous labels and each individual's feature distribution is represented by a Gaussian, using the resolved labels. For the test image, the joint probability $P(\mathbf{p} = \mathbf{n}|\mathbf{f})$ is estimated using the features $\mathbf{f}$. Inference is performed on the test image to determine an MPE assignment for all faces in the image, a MAP assignment for each face, an assignment of ambiguous labels, and the probability that each individual from that image collection is present in the image. It should be stressed that in this evaluation, each ambiguously labeled image contains at least two people, so the entire system works without a single face ever being positively identified by a user. The goal is to show classification is improved with the group prior.

The results are shown in Figures 8.8 - 8.11. Figure 8.8 shows the results for MPE, where the performance is the percentage of test images that all faces were correctly identified as a function of the amount of ambiguously labeled data. Set 2 proves to be the most difficult because individuals from this image collection have a large amount of overlap in the feature space. Figure 8.9 shows the results for MAP, where the classification rate is the percentage of faces that were correctly classified. Figure 8.10 shows the results for ambiguously labeling the test images, where the classification rate is the number of images that are assigned the correct ambiguous labels. Four different priors were used in each experiment. The group prior is the full model that includes both the unique object constraint and the prior for specific groups of individuals. The UOC prior enforces the unique object constraint, but assumes that each group of individuals has equal probability of appearing in an image (we use $P_{\text{UOC}}(n^u, n^v) \propto \alpha(u, v)$, from (8.8)). The individual prior ("Indiv") considers only the prior probability of an individual appearing in an image, and finally no prior at all is used ("none"). When using the individual prior or no prior, each face is classified as if it were the only face in an image, according to (8.2). Inference using the group prior and the UOC prior considers the features of all faces in an image for inference. By representing the social relationships between individuals with the group prior, the performance is nearly always improved over the UOC prior, sometimes by as much as 10-15%.

Figure 8.11 shows the accuracy of using the system to produce the score $P(n^q|\mathbf{f})$ that would be useful for an image retrieval system. The performance using the resolved ambiguous labels

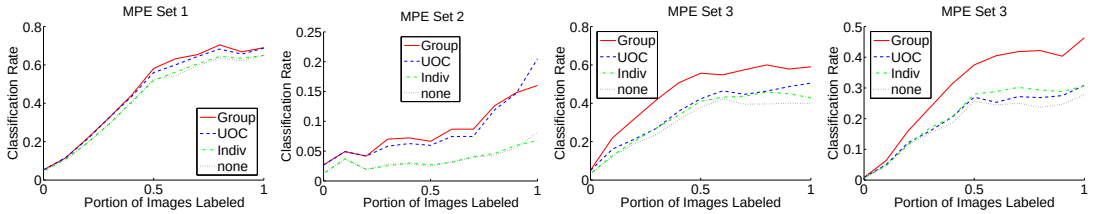


FIGURE 8.8: MPE performance on four consumer image collections using four different priors, as a function of the portion of the image collection with ambiguous labels.

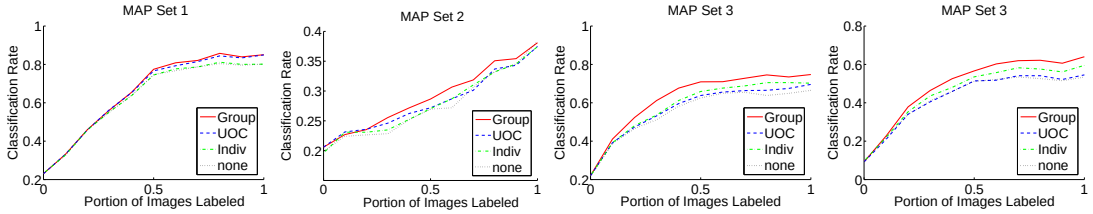


FIGURE 8.9: MAP performance on the four image collections.

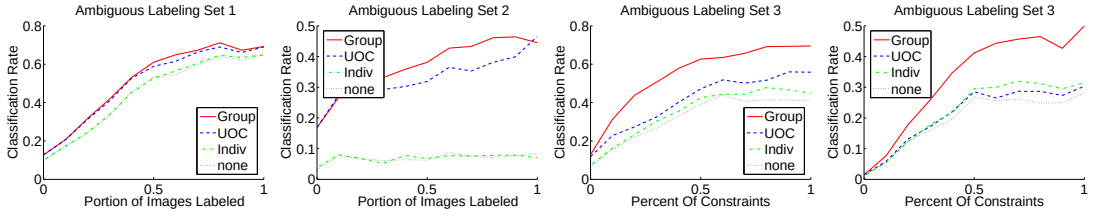


FIGURE 8.10: Performance of ambiguous labeling on the four image collections.

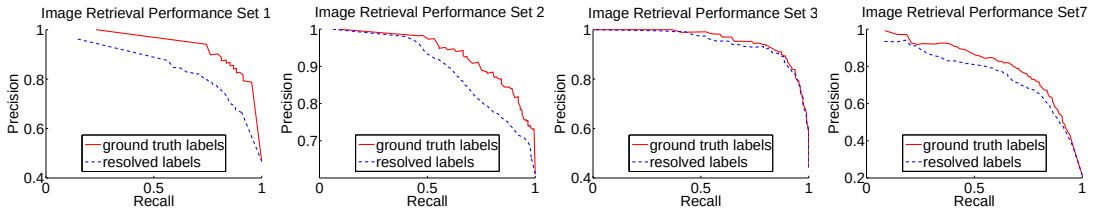


FIGURE 8.11: Retrieval performance on the four consumer image collections. In both cases the group prior is used. Training with resolved labels, which contain some mistakes, hurts the performance but the results are still very good.

is compared against using the actual ground truth labels, which is the upper bound for the performance of the ALR algorithm. The score $P(n^q|\mathbf{f})$ is produced for each test image for each individual in the set $\mathbf{N}$. Precision-recall curves are generated by varying a threshold on $P(n^q|\mathbf{f})$. All images except the randomly selected test image are ambiguously labeled. Mistakes made in resolving the ambiguous labels hurt the performance, but the recognition rates are surprisingly good, again considering that not a single face was explicitly labeled with the correct name.

FIGURE 8.12: An example of automatically resolved ambiguous labels for 15 images. Only two images contain mistakes, the last image of the first row, and the fourth image in the second row.

8.5 Discussion

We have introduced the problem of ambiguously labeled images in the context of labeling people in consumer image collections. We described an algorithm for resolving the ambiguous labels. Using the ambiguous labels, we learn a group prior for classification of people in unlabeled images. The group prior enforces the unique object constraint that an individual can appear at most one time in an image and indicates the probability of specific groups of people appearing together in an image. We demonstrated that despite errors in resolving ambiguous labels, useful classifiers can be trained with the resolved labels. By modeling the relationships between people in an image with the group prior, classification performance is significantly improved in all of our test sets.

Chapter 9

Multiple Contextual Features

In the preceding two chapters, we described several contextual features for identifying people in consumer images. In this chapter, we propose additional contextual features and describe a general framework for merging multiple contextual features in consumer image collections. While the number of possible contextual features that could be used is bounded only by our imagination, we are inspired by the contextual features that studies show are used by humans performing the same recognition tasks (see Chapter 2.4).

We model the instances of the faces in an image with a probabilistic graphical model, where each face is a node having a value from the set of all individuals in the image collection. A random subset of the images are labeled, and the model for appearance and context is learned from this subset. After learning, the model is applied to infer the identities of unknown faces in the collection. In this chapter, we show the results of using multiple contextual features in a unified model for several image collections.

The main contribution of this chapter is the unified model for merging context from any number of sources. In Chapter 2, a review of other approaches for incorporating context in the recognition of people is provided. Perhaps the most thorough exploration of multiple contextual features in person recognition is performed by O'Hare [97], where context (including clothing and geo-location) is incorporated with a hierarchical weighting of context and content features. However, this model does not maximize the likelihood of an assignment of identities to multiple people in an image, making it possible that multiple people in one image are all believed to be the same person. Our goal is to perform an assignment for the names of all faces in the image that considers all of the content and context with our model.

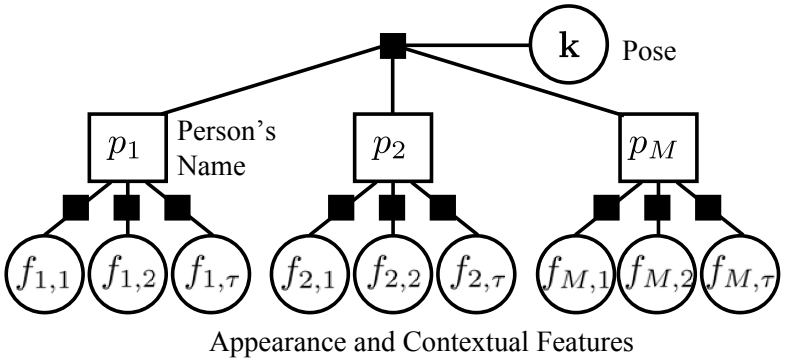


FIGURE 9.1: A factor graph representation of the model to combine appearance and contextual features for recognizing the identities of the M people in an image. Observed variables are shown with a circle, and hidden variables with a square. A solid square represents a factor in the model comprised of the variables that are connected to it. Features are assumed to be independent given the identity of the individual.

9.1 A Unified Contextual Model for Inferring Identity

In this section, we introduce the variables and the model that represents the identities and features in an image. 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 $\mathbf{N}$ (the set of all individual identities in an image collection) 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. Each person in the image has associated with it as set of c features. The feature $f_{m\tau}$ is the $\tau^{mathttth}$ feature vector for the $m^{mathttth}$ person in the image. The variable $\mathbf{k}$ represents the spatial arrangement of the faces in the image.

Without making any assumptions, the conditional distribution that the people $\mathbf{p}$ in an image have the identities $\mathbf{n}$ can be written:

$$P(\mathbf{p} = \mathbf{n}|\mathbf{f}, \mathbf{k}) = \frac{P(\mathbf{p} = \mathbf{n}|\mathbf{k})P(\mathbf{f}, \mathbf{k}|\mathbf{p} = \mathbf{n})}{P(\mathbf{f}, \mathbf{k})} \tag{9.1}$$

The terms of this model are:

$P(\mathbf{p} = \mathbf{n}|\mathbf{k})$ is the pose-dependent group prior. Given a particular arrangement of faces $\mathbf{k}$ in an image, this term is the belief that a particular explanation for all of their identities $\mathbf{p} = \mathbf{n}$ is likely. Note that this term contains all the information of the group prior from Chapter 8.

$P(\mathbf{f}, \mathbf{k}|\mathbf{p} = \mathbf{n})$ is the probability of observing the features and pose given a particular set of people are in the image, and $P(\mathbf{f}, \mathbf{k})$ is the probability of observing a particular set of evidence.

Name		Type	Dimension	Comment	Reference
Face	f_a	Appearance	61×49→37	Fisherface appearance feature	Chapter 7.2
Individual Prior		Social Context	1	Prior of individual appearance	
Clothing	f_c	Pixel Context	2×350	Bag of color and texture words	Chapter 7.2
Time	f_t	Capture Context	1	Used with clothing features	Chapter 7.5
Location	g_i	Capture Context	2	Latitude and Longitude	Chapter 9.3.3
Birthday	b_i	Social Context	1	Birthday of individual	Chapter 9.3.1
Group Prior	$P(\mathbf{p}=\mathbf{n})$	Social Context	$\binom{K}{M}M!$	Potentially Large	Chapter 8
Pose	$\mathbf{k}$	Pixel Context	2	$x-$ and $y-$ face position	Chapter 9.3.5

TABLE 9.1: A summary of the features considered by the unified model. Facial appearance (content) and contextual features and considered when determining a likely assignment for the faces in an image. For each feature, the feature dimension is noted.

FIGURE 9.2: After a face is resized to the standard 61×49 pixel size, it is projected into a subspace defined by 37 Faces learned from a separate training collection.

Figure 9.1 shows a factor graph representation of our unified model that represents the relationships between the identities of the people in the image and the observed content and contextual features. According to this model, the joint probability $P(\mathbf{p}=\mathbf{n}|\mathbf{f})$ of all the M people in a particular image, given the set of features $\{\mathbf{f},\mathbf{k}\}$ is written:

$$P(\mathbf{p}=\mathbf{n}|\mathbf{f},\mathbf{k}) = \frac{1}{Z}\Phi(\mathbf{p},\mathbf{k})\prod_{m,\tau}\Psi_{m,\tau}(f_{m,\tau},p_m) \tag{9.2}$$

The first term, $\Phi(\mathbf{p},\mathbf{k})$, is the pose dependent group prior. The second term $\Psi_{m,\tau}(f_{m,\tau},p_m)$ comprises the unary terms (one for each contextual feature for each person in the image). Each of these unary factors is of size $N \times 1$, where the $i^{\mathtt{th}}$ value relates to the relative likelihood that person m in the image has the $i^{\mathtt{th}}$ identity. Consistent with the model, the conditional distribution of the identities of people in the image is the product of the pose dependent group prior and factors related to each feature observation for each person in the image. The model incorporates several conditional independence assumptions that simplify the representation of the distribution. First, it is assumed that each person's appearance depends only on the identity of that person. The content and contextual features for a person do not change based on other people who are in the image. Second, the features associated with a particular person are conditionally independent given the identity of the person. This assumption is commonly made (i.e. the Naïve Bayes assumption) to simplify otherwise complicated distributions.

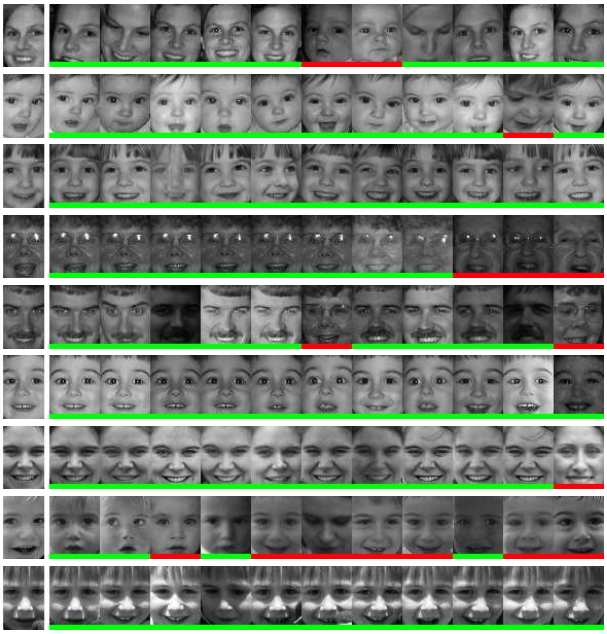


FIGURE 9.3: After projecting to a Fisher subspace, the nearest neighbors to a face of unknown identity (the leftmost image in each row) are found. Neighbors are underlined in green if their identity matches that of the query face and red if the identity does not match. From the neighbors, an estimate of $P(p = n|f_f)$ is produced.

9.2 Appearance Features

As in Chapter 7, features are extracted from the faces and clothing of people. Our implementation of a face detection algorithm [67] detects faces, and also estimates the eye positions. Each face is normalized in scale (61×49 pixels) and projected onto a set of Fisherfaces [11], representing each face as a 37-dimensional vector. Figure 9.2 shows a representation of the 37 Fisherfaces used in the projection.

The appearance term of the model $\Psi_f(f_f, p) = P(p = n|f_f)$ is the probability that a face p with the observes appearance belongs to a particular person n. This term is estimated with a Kernel Density estimate from the nearest neighbors. For a given face p, its nearest neighbors (in these experiments 9 nearest neighbors are used) from the set of labeled faces are found, where the distance is computed in the Fisher space. The term $P(p = n|f_c)$ is then estimated based on Maximal Likelihood Estimation using pseudocounts of the identities of the nearest neighbors. Figure 9.3 shows the nearest neighbors found for several faces from two different image collections.

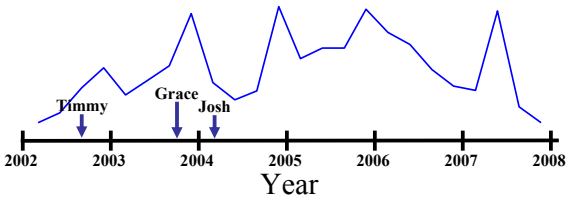


FIGURE 9.4: The lifespan of a person provides useful context for recognition. The timeline shows a histogram of images captured in one test collection between the years 2002 and 2008. During this span, three prevalent individuals in the collection are born. The people in an image from early 2002 cannot be either Timmy, Josh, or Grace.

9.3 Contextual Features

The model incorporates factors related to contextual features. The facial appearance and contextual features are described in Table 9.1. The following sections describe the features and factors from the model in more detail. When clothing and facial features are considered by the model, the fusion is performed at the feature level (a single factor $\psi_{fc}(f_{fc}, p)$ represents the joint relationship between contextual clothing features, facial appearance and identity. For other features (e.g. geo-location), a separate factor incorporates the information provided by the feature.

9.3.1 Birthday as a Feature

By simply knowing the birthdates of persons in the image collections as context, the recognition performance is improved as follows: Let the birthday of individual n_i of the set of identities be b_i. As described in Section 9.2, the nearest neighbors in the labeled training set are selected based on facial appearance (or, as shown in Section 9.3.2, a combination of clothing and facial appearance) are selected to estimate $\psi_f(f_f, p)$. When the birthday b_i of person is known, the nearest neighbors for an image of a query person are restricted to those from the set of all those training images having a birthday prior to the date that the image of the query person was captured.

In terms of the model, the factor $\Psi(b_i, t, p)$ depends on the birthdays of the persons in the set N as follows:

$$\Psi(b_i, t, p) = \begin{cases} 1, & b_i \leq t \\ 0, & b_i > t \end{cases} \tag{9.3}$$

where t is the time associated with the image capture time. Figure 9.4 illustrates the use of this contextual feature. In a similar fashion, the complete lifespans of the individuals in an image collection can be used as context.

9.3.2 Clothing Feature

Chapter 7 describes the clothing segmentation and clothing features in detail. To summarize, graph cuts are used to identify the clothing region of the image. Then, histograms of color and texture features are found to represent the clothing appearance as two histograms of 350 bins. Clothing similarity is found using the sum of χ^2 distances between the color and texture feature vectors. The clothing similarity is found considering the fact that people generally wear the same clothing throughout the duration of an event and is dependent on the time difference between the images. The appearance term of the model $\Psi_c(f_c, p) = P(p = n|f_c)$ is estimated using the identities of the nearest neighbors in a similar fashion to when the nearest neighbors are found from facial features.

When both clothing and face similarity are both considered by the model, the similarity between two persons is assumed to be based on facial similarity or clothing similarity, based on whichever feature provides the most similarity. Thus, the factor incorporating clothing context and facial appearance is expressed as:

$$\Psi_{fc}(f_{fc}, p) = P(p = n|f_{fc}) \tag{9.4}$$

The justification for combining face (content) and clothing (context) into a single factor of the model is that when either of these features presents strong evidence of similarity, it should be believed.

9.3.3 Geo-location

As observed by [35, 97], the location of an image is an important clue for inferring the identities of the people in the image. Intuitively, the people that are in the images captured by a photographer are somehow associated with a particular geographic location. For example, images

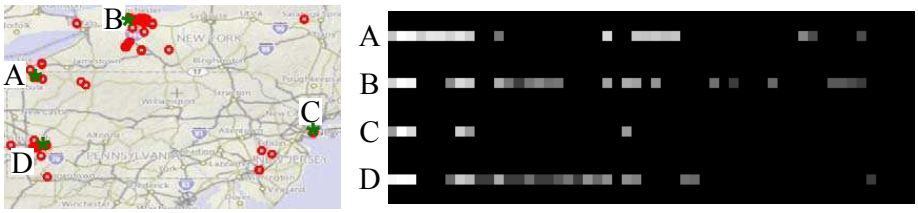


FIGURE 9.5: Red marks indicate the locations in the northeastern United States of images in the collection of 2990 geo-tagged consumer images. At four specific locations (roughly corresponding to Erie, Rochester, New York City, and Pittsburgh), the distribution over the 54 individuals in the collection is shown where light shades indicate higher likelihood of appearance.

captured at a friend's home are perhaps more likely to include that friend than others. The model contains a factor that captures the relationship between identity and geo-location.

An image is geo-tagged by either manually dropping images onto a map interface, or by using a bluetooth GPS device that communicates with a digital camera. Figure 9.5 shows the geo-locations of images in the geo-tagged image collection containing 2990 geo-tagged image where the images are captured in the Northeast of the United States, and a representation of the distributions of different individuals appearance at four specific locations.

In terms of the model, the factor $\Psi(g, p)$ is expressed as:

$$\Psi(g, p) = \left[\frac{P(p = n|g)}{P(p = n)}\right]^{\alpha_g} \tag{9.5}$$

where g represents the geo-location associated with the image. The distributions $P(p = n|g)$ and $P(p = n)$ are learned from empirical counts of the labeled portion of the image collection using a weak Dirichlet prior equal to the marginal prior distribution of an individual appearing in the collection as follows:

$$P(p = n|g) \propto \beta P(p = n) \sum_j \gamma_j \tag{9.6}$$

where β is a small constant, and

$$\gamma_j = \begin{cases} 1, & \texttt{if} n_j = n \texttt{and} \langle g_j, g \rangle < D_g \\ 0, & \texttt{else} \end{cases} \tag{9.7}$$

The distance between the geo-locations of the image in question (g) and the location of the $j^{\texttt{th}}$ person from the labeled portion of the image collection is $\langle g_j, g \rangle$ and is measured in kilometers. In (9.7), the distance threshold that is used is $D_g = 1$ kilometer.

The exponent α_g in (9.5) weights the influences of this factor relative to the other contextual features. In all our experiments, a training set is used to establish $\alpha_g = 0.4$, though we find that the results are not especially sensitive to this parameter, and this value is used for all collections in the test.

9.3.4 Group Prior and Position

The model includes the factor $\Phi(\mathbf{p}, \mathbf{k})$, a term that describes the relationship between the specific people in an image and their spatial relationships $\mathbf{k}$. This factor encompasses both the group prior from Chapter 8, as well as the spatial positions of the people in the image (first introduced in Chapter 3.

When the spatial position of people in the image is ignored, the factor becomes the group prior:

$$\Phi(\mathbf{p}) = \frac{P(\mathbf{p})}{\prod_m P(p_m)} \tag{9.8}$$

This factor represents the likelihood of any particular assignment of identities $\mathbf{n}$ to the people in the image. In Chapter 8, the group prior is represented as a product of pairwise potentials. As noted in 8, the complexity of the pairwise model grows exponentially with the number of people in the collection. For example, for an image with 6 people from an image collection with 54 distinct individuals, more than 18 billion assignments of names to faces need to be evaluated. In this unified model, we take a slightly different approach in order to simplify issues related to performing inference. This approach is at worst linear in the number of images in the image collection.

Instead of using a pairwise model, the labeled subset of the image collection is used to learn $P(\mathbf{p} = \mathbf{n})$. This term represents the distribution over all distict groups of people that are represented in the labeled portion of the image collection. The distribution has at most only j non-zero terms, where j is the number of images in the labeled training set. However, in practice, the number of non-zero terms is far fewer than that because many images capture the same group of people (either multiple images at the same event, or images of the same group at

Number People	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number Images	2323	852	179	52	21	6	3	0	0	0	2	2	2	4	2
Unique Groups	41	94	58	23	8	4	3	0	0	0	2	2	2	2	1

TABLE 9.2: In a collection of 3448 images containing 5092 instances of 54 unique people, there are 240 distinct groups. When inference is performed to determine an assignment of identities to faces in the image, this subset of groups (instead of every possible assignment) is examined.

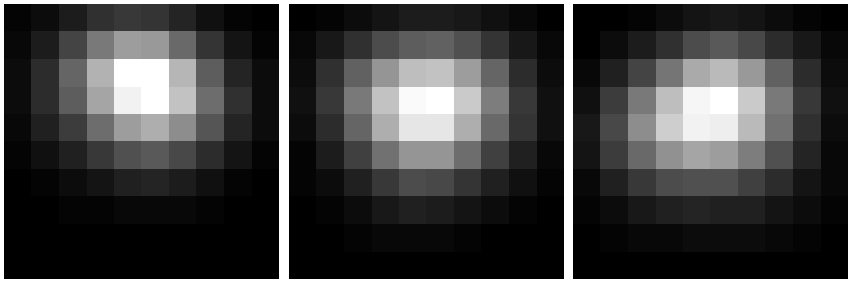


FIGURE 9.6: Because of the way they position themselves relative to others in the image, people appear at different positions in an image frame. Here, the distributions of $P(k_m|p_m = n_m)$ are illustrated for three different individuals in an image collection. The person represented on the left is the mother of the two children whose distributions are in the middle and on the right.

different events such a family portrait). Table 9.2 reports a summary of the distribution from an image collection of 3448 images, where there are only 240 distinct groups of people that appear in an image. Therefore, at inference, only sets $\mathbf{n}$ that are group subsets found from the labeled training subset are examined by the model. Although this approximation has no guarantees of finding global optima, the computational savings are a significant advantage.

9.3.5 Position as Context

In Chapters 3 and 5, the relative positions between people is a feature that provides information about the age, gender, and the relationship of the persons in the image. Here, we use the positions of faces as a feature that contains information about identity. It may seem counter-intuitive that considering facial positions can contribute to recognition. However, both physical and social factors result in non-random juxtapositions between people in an image. For example, a particular husband and wife may be photographed more often with the husband on the right and the wife on the left, although for another couple, the relative positions may be the opposite.

The positions of the faces of the people in an image are represented as $\mathbf{k}$. When position $\mathbf{k}$ is considered by the model, the group factor becomes:

$$\Phi(\mathbf{p}, \mathbf{k}) = \frac{P(\mathbf{k}|\mathbf{p})P(\mathbf{p})}{\prod_m P(p_m)} \tag{9.9}$$

where $P(\mathbf{k}|\mathbf{p})$ is the probability of a particular spatial arrangement given an assignment of identities to the M people in the image, $P(\mathbf{p})$ is the group group that represents the likelihood of a specific group of people appearing in an image, and $P(p_m)$ is the prior probability of person m appearing in a image.

Clearly, the more difficult term to deal with is the position-dependent group prior $P(\mathbf{k}|\mathbf{p})$. Just how can pose be represented so that this term can be represented? Even representing the distribution over $\mathbf{k}$, the positions of all the people in the image, is non-trivial especially when there are many (e.g. more than 2) people in the image. Even if a pairwise model is implemented to learn the relative pose between each pair of individuals in the image collection, the amount of training data is often insufficient. Some people appear in the collection only a limited number of times, and learning the high-dimensional pairwise model from limited data is difficult. Furthermore, even if the pairwise model is known, performing inference with it is difficult. Message-passing algorithms such as loopy belief propagation [84] can be used, but have no performance guarantees. In experiments, models that incorporated these more complex distributions for $\Phi(\mathbf{p}, \mathbf{k})$ are outperformed by the simplified representation we introduce next.

In our approach, we use a simplification of the position term $\mathbf{k}$ by learning the distribution over position in the image of each individual in the collection independently. While this simplification sacrifices some of possible benefits of learning a more complex model of position, it enables a simpler approach to learning and performing inference with guarantees. Consequently, the factor $\Phi(\mathbf{p}, \mathbf{k})$ becomes:

$$\Phi(\mathbf{p}, \mathbf{k}) = \frac{\prod_m P(k_m|p_m)P(\mathbf{p})}{\prod_m P(p_m)} \tag{9.10}$$

where $P(p_m|k_m)$ is the probability of a particular assignment of identity to the $m^{\texttt{th}}$ person in the image, given that person's spatial position in the image, $P(\mathbf{p})$ is the group prior, and $P(p_m)$ is the prior probability of person m appearing in a image.

The distribution $P(k_m|p_m)$ is learned from the subset of an image collection that is labeled by counting for each individual in the collection, the location of each face in a quantized (10×10)

representation of the image coordinates. Each appearance is assumed to be a sample with noise, so the distribution of the "true" location of the face is assumed to be a Gaussian with standard deviation of 1.0. Figure 9.6 shows the learned distributions of $P(k_m|p_m)$ for a few individuals from the image collections, and the ratio of $P(k_m|p_m)$. In the next section, we show how this representation for the positions of faces in the image simplifies inference, so that finding the global optimum assignments can be found (under some assumptions).

9.4 Inference

Using the model that incorporates the context and the appearance features, the goal is to find the assignment of identities to the people in an image that maximizes the likelihood of observing the features. When only one person is in the image, this is a trivial task. Each possible identity assignment is evaluated, and the assignment that maximizes (9.2) is the inferred assignment n_1 for the person p_1.

When multiple people are present in an image, the situation is more complicated but the goal remains the same; to find the optimal assignment of names $\mathbf{n}$ to people in the image.

$$\underset{\mathbf{n}}{\operatorname{argmax}} P(\mathbf{p} = \mathbf{n}|\mathbf{f}, \mathbf{k}) = \underset{\mathbf{n}}{\operatorname{argmax}} \Phi(\mathbf{p}, \mathbf{k}) \prod_{m,\tau} \Psi_{m,\tau}(f_{m,\tau}, p_m) \tag{9.11}$$

$$= \underset{\mathbf{n}}{\operatorname{argmin}} - \log \Phi(\mathbf{p}, \mathbf{k}) - \sum_{m,\tau} \log \Psi_{m,\tau}(f_{m,\tau}, p_m) \tag{9.12}$$

Equation (9.12) indicates that the optimal assignment of identities to people in the image is the one the minimizes a cost function, where the costs are related to the log of the factors of the model. Fortunately, an efficient method exists to find the optimal assignment of identities to people in the image $\mathbf{n}$, under the assumption that the image contains a group of people that has been seen in the training set. Borrowing from the insight used to infer with the group prior, we evaluate the model only for all the groups that are observed in the training set.

For each unique group of people in the training set, we find the optimal assignment for $\mathbf{n}$ as follows. Given a set of people in the image, we are searching for the permutation of the identities contained in $\mathbf{n}$ that optimizes (9.12). Because the term $P(\mathbf{P})$ is fixed for all assignments that are a permutation of the same set $\mathbf{n}$, this term can be ignored. Then, the assignment is the one that minimizes the sum of all the costs induced by all the feature observations for all the people

	Set 1	Set 2	Set 3
Total images	1065	5754	5359
Images with people	589	3448	2570
Number faces	931	5092	4502
Time span (days)	233	2361	1909
Unique individuals	32	54	34
Birthday?	No	Yes	Yes
Number Geo-tagged	0	2990	0

TABLE 9.3: A summary of the image collections used in the experiments.

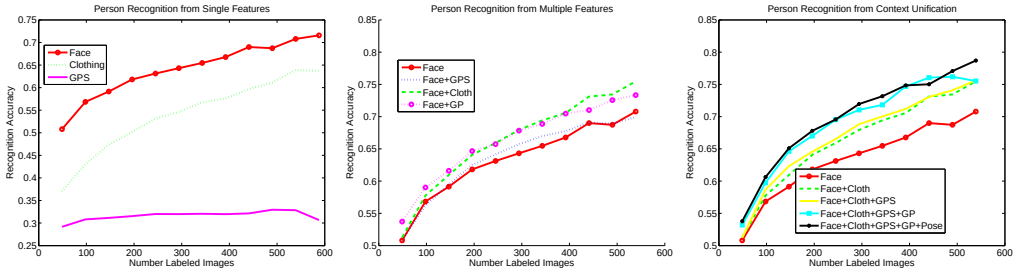
in the image $-\sum_{m,\tau} \log \Psi_{m,\tau}(f_{m,\tau}, p_m)$. This problem is the well-known assignment problem that is solved by Munkres algorithm [87].

To summarize, the optimal assignment of identities to the persons in an image is found by establishing the best possible assignment of labels to people for each unique group observed in the training set. For that optimal group, Equation (9.12) is evaluated. This process is repeated for all groups, and the group with the optimal value of Equation (9.12) is the optimal assignment, given the model parameters and the assumption that the image must be labeled with a group that has been observed in the training set rather than a novel group of identities. At worst, the complexity of the inference is $O(m^3 j)$ where m is the number of people in the image and j is the number of images in the (training) collection.

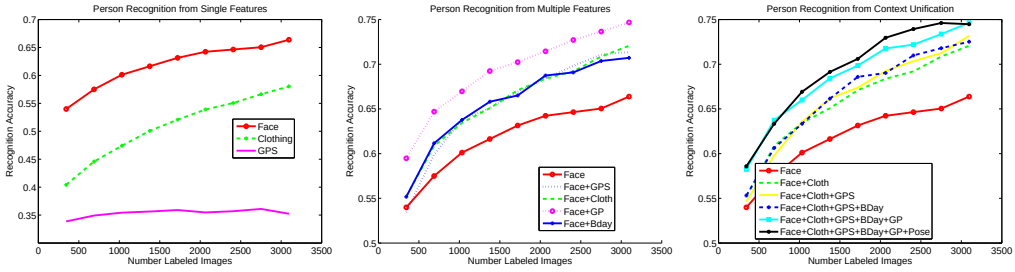
9.5 Experiments

To explore the performance of the model, experiments were performed on three image collections. Table 9.3 reports on the characteristics of the image collections. The image collections represent typical family photography, where the core family members appear most often in the collection although a set of friends and relatives are present though not as frequent.

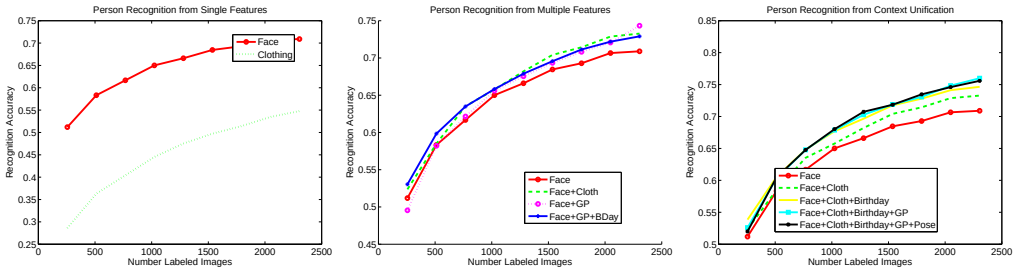
The testing procedure is as follows, independently on each image collection. A random subset of the images are selected and the true identities of the people in these images are provided for learning the factors of the model. The identities of all people in the non-selected images are inferred using the model, using some subset of the available contextual and appearance features. The inferred results are then compared with the actual identities to determine the accuracy of the model. This is repeated multiple times (at least 20) to generate a single point on the performance curve.



(a) Set 1



(b) Set 2



(c) Set 3

FIGURE 9.7: Context improves recognition in three consumer image collections, each row shows the performance of the model on a different collection. **Left:** The performance of individual features such as face, clothing, or geo-location. **Middle:** When individual contextual features are combined with facial appearance, the performance of the model improves. **Right:** As more contextual features are added to the model, the performance continues to improve, with the best performance being achieved when all context is considered.

Figure 9.7 shows the results of applying our model on several image collections. By using the unified model, recognition is improved as context is added. Each performance curve represents between 200,000 and 500,000 individual recognition events, and as a result the confidence interval on the curves is small, about 0.5%. The benefit of the contextual features is not equal across the collections. In these test collections, adding clothing features generally produces a large benefit, but adding geo-location results in only a modest improvement. In all the collections, the group prior provides an additional improvement, but considering the relative position of the faces in addition results in only a modest benefit (about 1% versus using the group prior

without position information). Set 3 shows the clear benefit of considering the birthday of the individuals in the collection. The important observation is that by considering as much contextual information as possible, the recognition of people in new images is improved.

9.6 Conclusions

In this chapter, we introduce a unified model for combining multiple contextual features with facial appearance for recognizing people in consumer image collections. In addition, we introduce several new contextual features. By considering the birthdays of those in the training set, we can prevent assigning inconsistent names when inferring the identity of people in an image. Further, we extend the group prior (a prior that expresses the association between groups of people) by considering the positions of people within the image. Due to physical and social factors, the position of people within the frame of an image is not completely random; but instead provides information about their identities.

Chapter 10

Conclusion

In this dissertation, novel contextual features are exploited to improve our understanding of images of people. Throughout the work on this dissertation, it became clear that there were many more related avenues to explore. In this chapter, we discuss future extensions of the work as well as a summary.

10.1 Future Direction

The majority of consumer images contain faces or people, and it is clear that understanding images of people will remain a primary goal of computer visionists for years to come. The remaining work is almost limitless. Modern computer vision can only answer questions about an image that are narrow in scope. Currently, most consumer image collections span only a few years to a decade (when digital camera quality began to rival that of traditional film systems). However, for many babies being born today, their entire lives will be captured digitally with cameras that record the time and place of the image capture. Throughout their livespan, they will each be the subject of tens or hundreds of thousands of images and videos. Managing and inferring the changing appearances and relationships that occur at this time scale will be a huge challenge.

Technically, many challenges remain to completely understanding images. Even when a face has been found, performing a good segmentation of the body of the person remains an open problem, though good progress has been achieved [78, 86, 104]. Then, the pose and body shape, size and proportions of the individual can be properly considered when making inferences

about identity, age, gender, and other demographic factors. Similarly, a complete system would incorporate content from features such as voice, gait, and person-specific facial expressions, and the contextual cues from 3D scene geometry.

Using context for the benefit of computer vision is the primary focus of this thesis. However, one could as easily reverse the question: How can computer vision be used to advance what we know about humans and their interactions? In one example of this idea, we showed in Chapter 3 that people stand, on average 306 mm from the closest neighbor in an image. As more and more images document nearly every moment of our lives, it becomes possible to use algorithms to answer such questions as: How close do people stand to one another in various social settings for various societies? How often do people socialize in mixed-gender versus single-gender groups? What percent of pairs of friends on social network sites actually spent time in the same physical location at some point in the past year?

10.2 Closing Summary

In this work, we extend the current knowledge in computer vision for images of people by showing unique contextual features for understanding images of people whether each image is part of a collection or exists as a single image. Perhaps the most important contribution of the thesis is the idea and corresponding evidence that the research results from the social sciences can be applied directly as social context to improve image understanding of images of people. The broad and intuitive idea is that the more we learn about people, the better we can interpret images of people.

Bibliography

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

[4] Apple. iphoto '09. Apr. 2009. http://www.apple.com/ilife/iphoto/.

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

[6] S. Bagon. Matlab wrapper for graph cut. Downloaded July 2007 from the Weizmann Institute. http://www.wisdom.weizmann.ac.il/~bagon.

[7] A. Balăn and M. Black. The naked truth: Estimating body shape under clothing. In *Proc. ECCV*, 2008.

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

[9] M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 2004.

[10] M. Behrmann, G. Avidan, J. Marotta, and R. Kimchi. Detailed exploration of face-related processing in congenital prosopagnosia: 1. Behavioral findings. *Journ. of Cognitive Neuroscience*, 2005.

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

[12] C. BenAbdelkader and Y. Yacoob. *Statistical Estimation of Human Anthropometry from a Single Uncalibrated Image*. Springer Press, 2008.

[13] 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.

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

[15] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene perception: detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982.

[16] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE PAMI*, 2003.

[17] W. Bledsoe. Man-machine facial recognition. Technical report, Panoramic Research Inc., 1966.

[18] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 2004.

[19] Y. Boykov, O. Veksler, and R. Zabih. Efficient approximate energy minimization via graph cuts. *PAMI*, 2001.

[20] V. Bruce and T. Valentine. Semantic priming of familiar faces. *Quarterly Journal of Experimental Psychology*, 1986.

[21] V. Bruce and A. Young. Understanding face recognition. *British Journal of Psychology*, 1986.

[22] L. Cao, M. Dikmen, Y. Fu, and T. S. Huang. Gender recognition from body. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, 2008.

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

[24] 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.

[25] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu. Composite templates for cloth modeling and sketching. In *Proc. CVPR*, 2006.

[26] L. Chen, B. Hu, L. Zhang, M. Li, and H. Zhang. Face annotation for family photo album management. *IJIG*, 2003.

[27] J. Y. Choi, S. Yang, Y. M. Ro, and K. N. Plataniotis. Face annotation for personal photos using context-assisted face recognition. In *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 44–51, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-312-9.

[28] R. Chopra and R. Srihari. Control structures for incorporating picture-specific context in image interpretation. In *Proc. IJCAI*, 1995.

[29] I. Cohen, A. Garg, and T. Huang. Vision-based overhead view person recognition. In *ICPR*, page 5119, 2000.

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

[31] T. F. Cootes and C. J. Taylor. Active appearance models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 484–498. Springer, 1998.

[32] A. Criminisi. *Accurate Visual Metrology from Single and Multiple Uncalibrated Images*. PhD thesis, University of Oxford, 1999.

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

[34] J. Cui, F. Wen, R. Xiao, Y. Tian, and X. Tang. Easyalbum: an interactive photo annotation system based on face clustering and re-ranking. In *Proc. ACM CHI*, 2007.

[35] M. Davis, M. Smith, D. Canny, N. Good, S. King, and R. Janakiraman. Toward context-aware face recognition. In *Proc. ACM MM*, 2005.

[36] S. K. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *Proc. CVPR*, 2009.

[37] B. Duchaine and K. Nakayama. Developmental prosopagnosia and the Benton facial recognition test. *Neurology*, 2004.

[38] B. Duchaine and K. Nakayama. Developmental prosopagnosia: a window to content-specific face processing. *Current Opinion in Neurobiology*, 2006.

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

[40] C. Dupertuis and J. Hadden. On the reconstruction of stature from long bones. *American Journal of Physical Anthropology*, 1951.

[41] M. Everingham and A.Zisserman. Automated person identification in video. In *Proc. CIVR*, 2004.

[42] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy – automatic naming of characters in tv video. In *Proc. BMVC*, 2006.

[43] M. Farah, K. Wilson, M. Drain, and J. Tanaka. What is "special" about face perception. *Psychological Review*, 105(3):482–498, 2008.

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

[45] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2), 2004.

[46] A. Fitzgibbon and A. Zisserman. On affine invariant clustering and automatic cast listing in movies, 2002.

[47] A. Gallagher. Consumer image person recognition database. http://amp.ece.cmu.edu/downloads.htm.

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

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

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

[51] A. Gallagher and T. Chen. Finding rows of people in group images. In *Proc. ICME*, 2009.

[52] A. Gallagher and T. Chen. Jointly estimating demographics and height with a calibrated camera. In *In Submission*, 2009.

[53] A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. CVPR*, 2009.

[54] A. Gallagher and T. Chen. A collection of group shots and row labelings. http://www.amp.ece.cmu.edu/people/andy/imagesOfGroups.html, by May 2009.

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

[56] 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.

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

[58] B. Golomb, D. Lawrence, and T. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Proc. NIPS*, 1990.

[59] Google. Picasa 3. Apr. 2009. http://picasa.google.com/.

[60] C. Gordon, B. Bradtmiller, T. Churchill, C. Clauser, J. McConville, I. Tebbetts, and R. Walker. 1988 anthropometric survey of US army personnel: Methods and summary statistics. *Technical Report NATICK/TR-89/044, AD A225 094*, 1988.

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

[62] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Automatic naming with caption-based supervision. In *Proc. CVPR*, 2008.

[63] 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.

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

[65] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

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

[67] M. Jones and P. Viola. Fast multiview face detector. In *Proc. CVPR*, 2003.

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

[69] T. Kanade. *Picture Processing by Computer Complex and Recognition of Human Faces*. PhD thesis, Kyoto Univ., 1973.

[70] I. Kispál and E. Jeges. Human height estimation using a calibrated camera. In *Proc. CVPR*, 2008.

[71] D. Klünder, M. Hähnel, and K.-F. Kraiss. Color and texture features for person recognition. In *IJCNN*, 2004.

[72] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *PAMI*, 2004.

[73] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *Proc. ICCV*, 2005.

[74] J.-F. Lalonde, D. Hoiem, A. A. Efros, C. Rother, J. Winn, and A. Criminisi. Photo clip art. *ACM Trans. SIGGRAPH*, 2007.

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

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

[77] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[78] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proc. CVPR*, 2005.

[79] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[80] D. Liu and T. Chen. Background cutout with automatic object discovery. In *Proc. ICIP*, 2007.

[81] 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.

[82] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):541–547, 2008.

[83] 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.

[84] T. Meltzer and Y. Weiss. c_inference. Downloaded Nov. 2006 from Hebrew University. http://www.cs.huji.ac.il/~talyam.

[85] Microsoft. Easyalbum demo software. Apr. 2009. http://research.microsoft.com/en-us/groups/vc/easyalbumdownload.aspx.

[86] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *Proc. CVPR*, 2004.

[87] J. Munkres. Algorithms for the assignment and transportation problems. *SIAM*, 1957.

[88] E. Murphy Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE PAMI*, 31(4):607–626, April 2009.

[89] M. Murty, A. Jain, and P. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 1999.

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

[91] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *Pattern Recognition*, 2003.

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

[93] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, 1998.

[94] A. Ng, Y. Weiss, and M. Jordan. On spectral clustering: Analysis and an algorithm. In *Proc. NIPS*, 2002.

[95] M. H. Nguyen, J.-F. Lalonde, A. A. Efros, and F. de la Torre. Image-based shaving. *Computer Graphics Forum Journal (Eurographics 2008)*, 27(2):627–635, 2008.

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

[97] N. O'Hare. *Semi-Automatic Person-Annotation in Context-Aware Personal Photo Collections*. PhD thesis, Dublin City University, 2007.

[98] N. O'Hare and A. Smeaton. Context-aware person identification in personal photo collections. *IEEE Trans MM*, 2009.

[99] S. Palmer. The effects of contextual scenes on the identification of objects. *Memory and Cognition*, 1975.

[100] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE PAMI*, 22(12):1424–1445, 2000.

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

[102] W. Pennebaker and J. Mitchell. *JPEG Still Image Data Compression Standard: Still Image Data Compression Standard*. Springer, 1993.

[103] 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.

[104] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proc. CVPR*, 2005.

[105] W. Rand. Objective criteria for the evaluation of clustering methods. *Journ. American Statistical Association*, 1971.

[106] C. Rother, V. Kolomogorov, and A. Blake. Grabcut- interactive foreground extraction using iterated graph cuts. In *Proc. ACM Siggraph*, 2004.

[107] 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.

[108] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proc. CVPR*, 2006.

[109] S. Satoh and T. Kanade. Name-it: Association of face and name in video. In *Proc. CVPR*, 1997.

[110] H. Schneiderman and T. F. A statistical method for 3d recognition applied to faces and cars. In *Proc. CVPR*, 2000.

[111] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 2000.

[112] 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.

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

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

[115] P. Sinha, B. J. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: 19 results all computer vision researchers should know about. *Proc. of the IEEE*, 2006.

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

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

[118] M. Smith, P. Morris, A. Collins, and P. Levy. *Cognition in Action*. Psychology Press, 1994.

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

[120] N. Sprague and J. Luo. Clothed people detection in still images. In *Proc. ICPR*, 2002.

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

[122] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *Proc. CVPR*, 2005.

[123] D. Thompson, S. Robertson, and R. Vogt. Person recognition: the effect of context. *Human Learning*, 1982.

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

[125] A. Torralba. Contextual priming for object detection. *IJCV*, 52(2):169–191, 2003.

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

[127] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 1991.

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

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

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

[131] A. Whiteley and E. Warrington. Prosopagnosia: a clinical, psychological, and anatomical study of three patients. *Journal of Neurolory, Neurosurgery, and Psychiatry*, 40:395–403, 1977.

[132] E. Winograd and N. Rivers-Bulkeley. Effects of changing context on remembering faces. *Journal of Experimental Psychology: Human Learning and Memory*, 1977.

[133] L. Wiskott, J. Fellous, N. Kruger, and C. V. D. Malsburg. Face recognition by elastic bunch graph matching. *IEEE PAMI*, 1997.

[134] L. Wolf and S. Bileschi. A critical view of context. *IJCV*, 2006.

[135] C. Wu, C. Liu, H.-Y. Shum, Y.-Q. Xy, and Z. Zhang. Automatic eyeglasses removal from face images. *IEEE PAMI*, 26(1):322 – 336, March 2004.

[136] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. http://www.cse.msu.edu/~yangliu1/frame_survey_v2.pdf, 2006.

[137] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE PAMI*, 2002.

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

[139] A. Young, D. Hay, and A. Ellis. The faces that launched a thousand slips: Everyday difficulties and errors in recognizing people. *British Journal of Psychology*, 1985.

[140] S. Yu, R. Gross, and J. Shi. Concurrent object recognition and segmentation by graph partitioning. In *Proc. NIPS*, 2002.

[141] S. Yu and J. Shi. Grouping with bias. Technical report, CMU, 2001.

[142] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *Int. J. Comput. Vision*, 8(2):99–111, 1992.

[143] 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.

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

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

[146] Z. Zhang. A flexible new technique for camera calibration. *IEEE PAMI*, 2001.

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

[148] X. Zou, J. Kittler, and K. Messer. Illumination invariant face recognition: A survey. In *IEEE Conf. on Biometrics: Theory, Applications, and Systems*, pages 1–8, 2007.