UNIVERSITY OF CALIFORNIA, SAN DIEGO

Two Faces are Better Than One -
Face Recognition in Group Photographs

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science

by

Ohil Krishnamurthy Manyam

Committee in charge:

Professor David Kriegman, Chair
Professor Serge Belongie
Professor Charles Elkan

2011

The thesis of Ohil Krishnamurthy Manyam is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California, San Diego

2011

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Prof. David Kriegman for two challenging and extremely enriching years of research. Through multiple ideas and many experiments - not all successful, his patience and timely guidance were largely responsible for bringing this work to its completion.

I would like to thank Prof. Peter Belhumeur and his graduate student Neeraj Kumar at Columbia University, New York for the many fruitful and helpful telephonic discussions. A special thanks to Neeraj for his help in generating attributes for the datasets in this work.

Thanks also to Florian Schroff at UCSD for helping me get up to speed with the Buffy dataset, and for valuable pointers to many face recognition tools along the way.

For their valuable suggestions that greatly helped improve the quality of this manuscript, I would like to thank my thesis committee members Prof. Charles Elkan and Prof. Serge Belongie.

Finally, thanks to my parents in India, and my relatives in the US for all their support ... and Indian food!

ABSTRACT OF THE THESIS

Two Faces are Better Than One -
Face Recognition in Group Photographs

by

Ohil Krishnamurthy Manyam

Master of Science in Computer Science

University of California, San Diego, 2011

Professor David Kriegman, Chair

Given an image containing more than one individual, face recognition systems so far have assumed statistical independence between each detected face when making a recognition decision. Contrary to this, for face recognition in unconstrained and natural settings, we show that there is potential for an increase in recognition accuracy by identifying people in groups. We propose models based on conditional and joint probabilities for handling recognition of pairs of individuals. These models are subsequently evaluated on two datasets - one from a television show and another, a personal photo album. In addition to using various state-of-the-art attribute based features, we design new descriptors of our own that can capture naturally occurring color and height correlations in group images. We

report recognition accuracy achieved by our relative models and compare this to existing models that assume statistical independence. We examine issues related to data scarcity when building relative models and propose techniques to combine group recognition decisions with statistical independence decisions to overcome these issues. Although improvements in accuracy over baseline techniques are modest for our implementation, we show that there is indeed potential in relative face recognition by using color and height based descriptors in conjunction with our relative models.

Chapter 1

Introduction to Face Recognition

Recognizing people in photographs comes rather naturally to humans. Machines on the other hand have to be taught - first to detect or localize a human face within a photograph, and then to actually recognize that person. Both these tasks have been researched extensively over the past few years. For face detection, several techniques exist [1], but perhaps the most popular approach among them all is the one by Viola and Jones [2]. This technique involves evaluating simple Haar wavelet based features at various scales and positions on an image. An integral image is used to speed up the computation of these features - even achieving real-time speeds for moderately sized images. The features are finally fed to a classifier cascade trained using AdaBoost. Face detectors are getting increasingly better at localizing faces even in the presence of variations due to lighting, pose, or image resolution.

Face recognition is the process of identifying a person from his or her image. Such a system is typically trained with a large number of characteristic images for every person that the system is designed to recognize. Using these images, a recognition model can be built that is either generative or discriminative. In a generative approach, the system builds a mathematical model corresponding to each person's facial appearance with the goal of describing the entire face image in detail. A discriminative approach on the other hand places greater importance on specific traits that help in discriminating people, rather than describing each individual. With either approach, the actual recognition task usually starts with a

face detector segmenting a specific region containing a face from an input image. This region is then fed to the face recognizer which tries to find a person whose trained model best describes the input region. The identity of such a person is finally assigned to the detected face region.

Various techniques have been proposed in literature to fulfill the task of building a model for each person. Historically, Eigenfaces [3] and Fisherfaces [4] have been used extensively. With improvements in image capture technology, 3D face models have been receiving increasing attention [5]. Considerable effort has gone into making recognition systems invariant to pose [6] and illumination changes [7], although a reliable system that can handle all such variations remains elusive.

Among existing systems, very high recognition accuracies ($\approx 100\%$) are common [8] when subjects are photographed over short timespans, in well lit environments, with known lighting color and directions, limited face pose variations and easily subtracted backgrounds. As one might expect, relatively simple techniques such as pixel-to-pixel comparisons also work well in this case owing to the many constraints. In practice, such a situation can occur in controlled environments with cooperative users. For example, a face recognition system that controls entry into a secure facility. A user in this case, can be expected to face the camera from an appropriate distance and maintain a frontal pose until he or she is identified.

There are many real life applications where the previously mentioned constraints are too narrow and strict. Personal photo albums for instance - where individuals may be photographed in a natural setting with a wide variety of pose, lighting, expression and image quality. With the advent of social networking sites such as Facebook, millions of such images are even uploaded daily. Automatic annotation of faces in photographs is a valuable tool for sites such as Facebook as well as photo album managing software like Google's Picasa, Apple's iPhoto or Microsoft's Windows Live Photo Gallery. A major hurdle to accurate face recognition in personal photo albums is the lack of constraints on input images.

1.1 Unconstrained Face Recognition

Of late, there has been an increased interest in relatively unconstrained face recognition problems. Databases such as Labeled Faces in the Wild (LFW) [9] and PubFig [10] for example, present images of celebrities for face recognition. These images are obtained from an image search on the internet - and so have very few image constraints. In addition, since the image acquisition is spread over long periods of time, there may even be changes in the physical appearance of a face. Unsurprisingly, many techniques that worked well for the constrained face recognition case perform rather poorly here. However, techniques have been proposed to deal specifically with this problem.

Guillaumin et.al [11] uses metric learning techniques including a logistic discriminant based approach and a nearest neighbor approach to perform face verification on the LFW dataset. They experiment with features based on Local Binary Patterns (LBP) and the Scale Invariant Feature Transform (SIFT). Cao et al. [12] demonstrate a descriptor that employs a learning-based encoding scheme on data extracted from image patches around fiducial points. Kumar et al. [10] proposed an attribute based framework wherein each individual is assigned a collection of scores from various attribute classifiers. In this work, we use the same attribute framework to generate feature vectors describing each candidate in our dataset. We use these feature vectors to build a discriminative model for recognition.

1.2 Relative Model for Face Recognition

Face recognition systems to date, despite using different techniques to generate descriptive feature vectors, end up building models that describe each individual separately and independently. Techniques that consider cues from regions outside the face box tend to model context [13–16] to help in recognition. These cues include clothing color and body feature descriptors still computed per-person. Beyond contextual cues, techniques that use metadata in the form of connectivity graphs from social networks were reported in [17, 18]. Gallagher et al. [19, 20]

developed models that correlated the position and height of multiple people in photographs to their age, gender and social relationship. Here too, metadata was modeled.

Figure 1.1: Collection of frames containing multiple individuals from the television show Buffy the Vampire Slayer

In this work, we approach the problem of recognizing faces in group photographs at a more grass-root level instead of modeling metadata. We explore various visual similarities and conditions that people share in a group photograph. Several such similarities can be seen in Figure 1.1. Due to a common light source, the direction of shadows cast are all the same and can possibly be ignored. Measured skin tones of different people can be directly compared within an image without worrying about lighting normalization or color correction. The same height relationships (taller-shorter) between people can be observed in multiple images. Also, the direction of gaze and pose of individuals in an image are highly correlated - everyone is staring at a common point of interest.

To exploit such similarities, we build relative models for pairs of individuals from automatically computed descriptors based purely on the visual image. Our system produces recognition decisions for pairs of individuals at once. To the best

of our knowledge, this is the first attempt at building such relative recognition models. Our approach represents a major departure from existing methods that use visual cues to recognize each person independently of the others. Techniques that work with social group data do so at a post-processing stage after candidate identification labels have already been generated per-person. In addition to recognizing people in pairs, unlike other techniques mentioned, our system models image characteristics rather than metadata. Thus, it is feasible to add-on our technique as a module to existing face recognition systems, providing an improvement in recognition rates for group photographs. Although we focus on recognizing pairs of individuals, the essential ideas we present can be used with larger groups.

Following this introductory chapter, Chapter 2 details our framework for generating feature vectors based on attributes. We also describe additional color and height based descriptors. In Chapter 3 we introduce the datasets that we use for our experiments. Chapter 4 explores naturally occurring commonalities in group photographs, presents our experiment framework and reports results for baseline experiments. In Chapter 5, we propose relative recognition models that handle pairs of individuals and also report accuracy results for the same. We explore techniques to fuse decisions from one of our relative models and the baseline model in Chapter 6. We conclude in Chapter 7 and hint at possible future work in Chapter 8.

Chapter 2

Features Vectors to Describe a Face

One of the first steps in any recognition task is to build a feature vector for each input that is able to capture characteristic traits which have high discriminative power. It is generally observed that more discriminative features lead to high recognition accuracy even with simple classifiers. In the spirit of this observation coupled with our intention to work with natural group photographs, we use features that have been shown to work well in an unconstrained setting. In addition, we include a few new descriptors of our own.

2.1 Attribute Features

Kumar et al. [10] introduced a novel approach to face verification using attribute detectors. As defined by them, attributes are describable aspects of visual appearance such as gender, race, age and hair color. Attributes are formulated such that they are either present or absent in an image. Thus, *Male* is used as an attribute, but not *Gender*. To automatically detect the presence of these attributes, Support Vector Machines (SVMs) employing an RBF kernel were used. The SVMs were trained with images over a wide range of conditions to make them invariant to pose, lighting, expression etc while still detecting specific attributes. In their work, Kumar et al. used outputs from each of 65 attribute detectors as a feature

vector to describe every individual. Such feature vectors from two individuals was fed to another SVM to obtain a face verification decision with 85.29% accuracy.

For our work using personal photo albums, the face recognition process needs to work with rather unconstrained inputs. Variations in pose, expression, lighting and image quality are common. Due to the demonstrated discriminative power of the attribute framework, we employ this as one set of features for our experiments. While we use attributes for face recognition, it may be noted that in case of Kumar et al. face verification was performed. Further, in our case, although the attribute detecting SVM outputs are intended as binary decisions, the raw values returned are retained and used as an indication of the *degree* of presence or absence of an attribute. SVMs in general are trained for maximum margin separation, and the raw output is a distance metric from the input point to the separating hyperplane. For example, when a male attribute detector trained to output $+1$ for male and -1 for female returns 0.9 as its output, this is seen as an indication that the male attribute is substantially dominant in the image being tested - i.e. the image is on the male side of the male-female separating SVM hyperplane and is far away from this hyperplane. Similarly, a decision value of -0.2 by the same male detector indicates a weak female attribute (closer to the hyperplane).

Table 2.1 presents a list of all 73 attributes used in this work. It may be noted that some attributes are highly discriminative between people (Asian, Brown Eyes) while others are not (Smiling, Outdoor). The weakly discriminative attributes are still important as they capture a general quality associated with a person. For instance, a person who tends to smile heavily in photographs would have a consistently high value for his/her Smiling attribute. Similarly a person who often wears a tie may have a high average value for his/her Wearing Necktie attribute. Values for all these attributes were computed by an online web-service maintained at Columbia University, New York. Although we used the attribute generator system as a black-box, the basic steps involved face detection, fiducial point detection, face alignment based on the fiducial points, followed finally by attribute detection.

Table 2.1: A list of 73 attributes that we use

5 o' Clock Shadow	Flash	Posed Photo
Arched Eyebrows	Flushed Face	Receding Hairline
Asian	Frowning	Rosy Cheeks
Attractive Man	Fully Visible Forehead	Round Face
Attractive Woman	Goatee	Round Jaw
Baby	Gray Hair	Senior
Bags Under Eyes	Harsh Lighting	Shiny Skin
Bald	Heavy Makeup	Sideburns
Bangs	High Cheekbones	Smiling
Big Lips	Indian	Soft Lighting
Big Nose	Male	Square Face
Black	Middle Aged	Straight Hair
Black Hair	Mouth Closed	Strong Nose-Mouth Lines
Blond Hair	Mouth Slightly Open	Sunglasses
Blurry	Mouth Wide Open	Teeth Not Visible
Brown Eyes	Mustache	Wavy Hair
Brown Hair	Narrow Eyes	Wearing Earrings
Bushy Eyebrows	No Beard	Wearing Hat
Child	No Eyewear	Wearing Lipstick
Chubby	Obstructed Forehead	Wearing Necklace
Color Photo	Outdoor	Wearing Necktie
Curly Hair	Oval Face	White
Double Chin	Pale Skin	Youth
Eyeglasses	Partially Visible Forehead	
Eyes Open	Pointy Nose	

2.2 Color and Height Based Features

Variations in color and lighting have long been considered a hindrance to accurate face recognition [21]. Many systems completely throw away color information by working purely with grayscale images. Such systems usually have a pre-processing stage where color and lighting intensity variations are normalized before being discarded. On the other hand, there have been many attempts which show improvements in accuracy when color is leveraged for recognition [22–24]. Despite this advantage, systems continue to ignore color information because of sufficient discriminative traits available in grayscale normalized images. For example, a system may be able to learn the shape of a face and positions of various parts such as eyes, nose and mouth within a face box. With good quality images, these alone may be sufficient for accurate recognition. But with low resolution images, color may very well become an important cue [25]. To illustrate this better, we carry out a preliminary test of our own.

2.2.1 A Preliminary Color Experiment

When a device (such as a camera) is used to capture an image of an object, it is well known that the image is influenced by the intensity, color and direction of incident light, the color and shape of the object and the spectral response of the image capturing device. To better understand the interplay of these parameters with respect to face recognition, we use a physics-based face database from The University of Oulu, Finland [26]. The database consists of 125 individuals photographed in a single session, with frontal pose, a fixed neutral expression and under four different lighting conditions - incandescent CIE illuminant A (Planckian 2856K), horizon daylight (Planckian 2300K), fluorescent TL84 (F11 4000K) and D65 (Daylight 6500K) [1]. Although the same camera was used throughout the database, it was white balanced to match all four lighting conditions for each for each incident lighting - thereby producing 16 images per person. Figure 2.1 shows the 16 images captured for one person. Images along the diagonal were captured

[1] These are respectively abbreviated as *a*, *h*, *t* and *d* in Figures 2.1 and 2.2

with the camera white balanced for the respective illumination. A drastic change in visual appearance can be seen across images. As expected, a simplistic face recognition experiment using 16×16 pixels downsampled versions of these images while retaining each of the three - red, green and blue - color channels is not very accurate. Using only images where the camera calibration and incident lighting matched (images along the diagonal in Figure 2.1), face recognition accuracy is in the 90% range. For example, training on images with incident light and camera calibration h and testing on images with incident light and camera calibration t provides 90.4% accuracy. But when the training and testing images do not have the same camera calibration and incident lighting, accuracies vary from 88.8% (train: camera h, light t; test: camera a, light t) to as low as 8.8% (train: camera h, light a; test: camera t, light h). This trend persisted with a few other color spaces as well [2].

To avoid similar drastic fluctuations in accuracy, color and lighting normalization techniques have been used as a pre-processing stage in many face recognition systems. Figure 2.2 was produced by first converting each color image from Figure 2.1 into a grayscale version. A popular equation $I = 0.2989R + 0.5870G + 0.1140B$ was used where I represents the intensity of a grayscale pixel and R, G and B represent the red, green and blue channel values of the color pixel respectively. This equation is used for instance by the NTSC television standard for computing the luminance component. Following the conversion to grayscale, histogram equalization was applied. Histogram equalization is a contrast adjustment technique which spreads an input image's grayscale histogram uniformly over the entire grayscale range. As can be seen, the face images in Figure 2.2 appear remarkably similar. Indeed, this translates to an increased accuracy ($\approx 99\%$) following the same recognition framework as before where a downsampled version of each image is used as its feature vector. The normalization is so helpful that high accuracy can be achieved irrespective of the imaging conditions for the training and test images. Effectively, changes in camera calliberation, color and intensity of lighting have all been nullified through this transformation.

[2]This experiment is meant to illustrate a possible detrimental effect due to the inclusion of color based features. It is not intended to be rigorous.

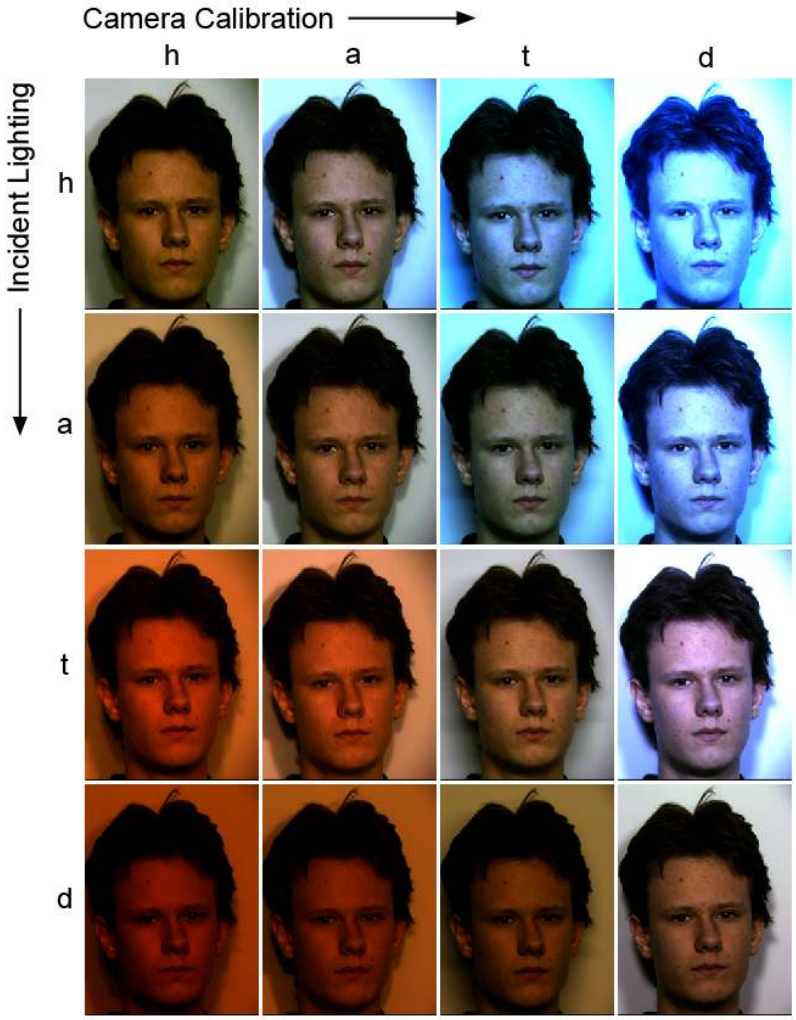

Figure 2.1: Original images for an individual from the Oulu physics-based face database [26]

When compared to their color versions, the histogram equalized grayscale images certainly provide better recognition accuracy. But it should be noted that various other factors within this dataset influence this increase in accuracy. Images for each individual were captured over a short timespan. In addition, there is very little pose variation per person. A fixed color background and uniform image resolution and quality are other factors that help maintain a high recognition accuracy for the grayscale versions. Since our goal is to work in the unconstrained face recognition setting, most of these factors will not remain constant across images

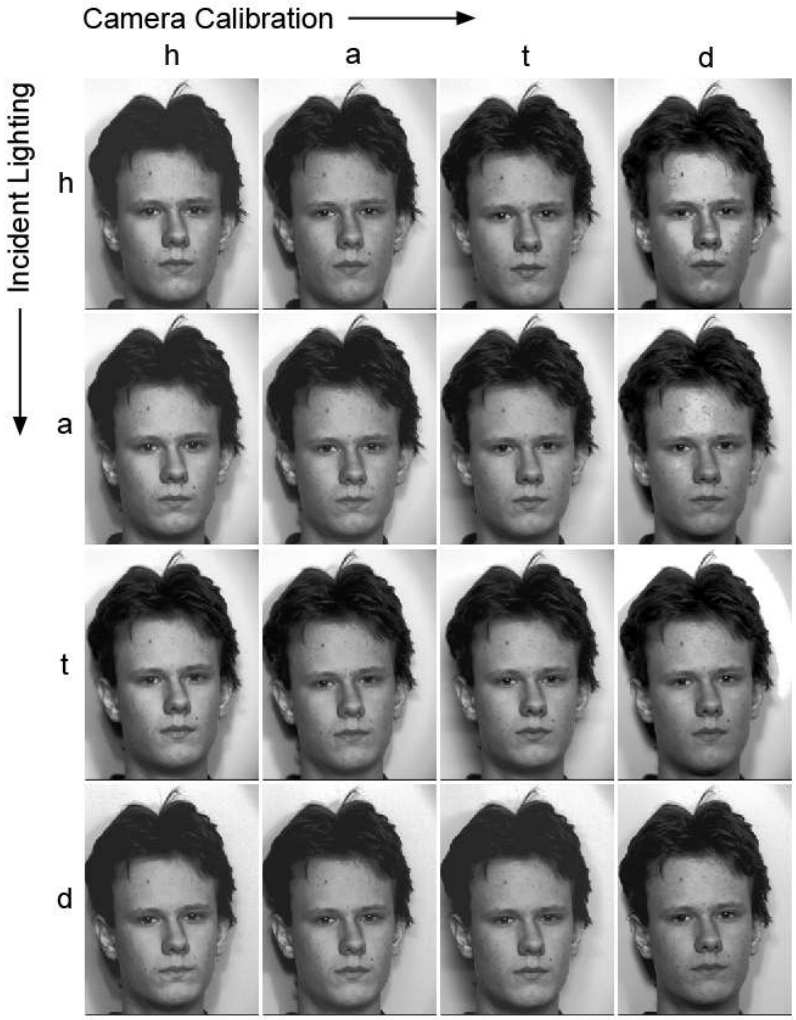

Figure 2.2: Grayscale histogram equalized images for an individual from the Oulu physics-based face database

in our case. In such a situation, color may very well be an important indicator. Figure 2.3 illustrates this point - while skin color can be a useful feature in color images, this information is almost completely lost in the normalized version.

2.2.2 Introducing New Descriptors

In addition to the attribute features mentioned, we introduce a few new descriptors of our own. These descriptors are based on color and height of an individual corresponding to a detected face box. Our reason for introducing these new

Figure 2.3: Color and histogram equalized images of two individuals in the Oulu physics-based face database

descriptors are two fold. First, considering that our recognition experiment is in a tough unconstrained setting, we hope to harness any extra information that color or height can provide. In addition, color and height measurements for face images of individuals are highly correlated when they occur in the same group photograph. For example, if two individuals are both seen in multiple photographs, their skin tones would be correlated across photographs due to common light sources and imagining conditions. We hope that our relative recognition models are able to learn these correlations and provide a boost in accuracy.

Color Descriptors

To capture characteristic color traits, we introduce four new descriptors corresponding to four regions within the rectangular boundary of a detected face. As mentioned previously, one stage in the attribute generation pipeline is fiducial point detection. This produces six fiducial points corresponding to the corners of

each eye and corners of the mouth. With this, we define four regions from which we extract color descriptors. Two regions correspond to rectangular areas around each eye. One region corresponds to a similar rectangular area enclosing the mouth. Each rectangular region contains two fiducial points, and the region itself is defined to have a width that is 115% of the distance between the two fiducial points and a height that is one half of this width. We define one final region as the convex hull enclosing all six fiducial points [3]. Figure 2.4 shows fiducial points detected on an image followed by regions that are used for the generation of our new color based feature vectors. It may be noted that additional regions can be defined in a similar way when more fiducial points are available - for instance around the forehead.

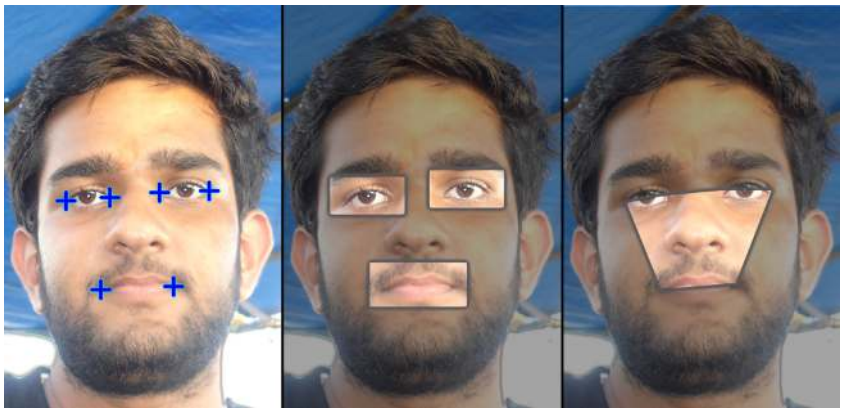

Figure 2.4: Fiducial points detected (left), eyes and mouth regions (center) and convex hull of fiducial points (right)

For generating color descriptors from each region defined above, we experimented with four color spaces (RGB, YCbCr, HSV and L*a*b*), seeking a single three component descriptor for each region. We considered mean as well as median descriptors where each color channel's mean (or median) is computed independently - resulting in a final three component feature per region. Cross validation revealed the HSV (Hue Saturation Value) median descriptors to be the best among those considered. Specifically, for the HSV (Hue Saturation Value) color space, while the saturation and value components vary linearly, the hue com-

[3]Initial experiments using the entire face box instead of the convex hull resulted in slightly lower cross validation accuracy. We conjecture that this is due to extraneous background colors showing up in the rectangular bounding box around the face.

ponent varies in an angular fashion. Red - which is predominant in skin color and consequently in all of our descriptor regions - occurs at 0° hue. Sometimes, this causes measured hue values to lie around 360°. In the interest of avoiding such jumps in measurements and to maintain a continuous variation in all our feature vector components, we enforce a hue shift of 180°. It was this hue-shifted HSV median descriptor that was found to perform best and so is chosen as the color descriptor for all four regions surrounding fiducial points on a face.

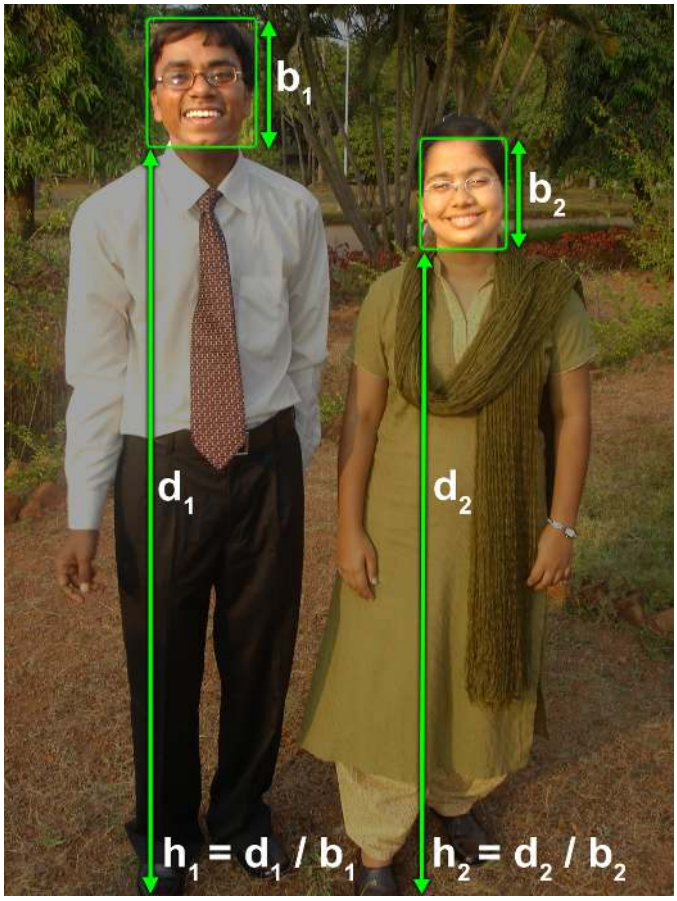

Figure 2.5: Computing height descriptors h_1 and h_2 for two people in a group photograph

Height Descriptor

Similar to color, the height of a person can also be a good relative descriptor. Due to variations in camera position and ground level, the height of a person as

estimated from a photograph may experience drastic fluctuations. Consequently, height may be a very weak descriptor when recognizing people in a group photograph with the statistical independence assumption between people. On the other hand, the face box of a taller person is more likely to be found higher up in a photograph than that of a shorter person. While it is true that this relationship can easily get disturbed (say when the shorter person wears high-heeled shoes), a probabilistic model can be expected to learn a relative height distribution.

Thus, in addition to the color descriptors, we introduce one final descriptor corresponding to the height of a person as seen in an image. The distance in pixels between the center of a detected face box and the base of the image is representative of a person's height. To account for people closer to the camera appearing larger and taller, we take a ratio of the distance of the face box from the base of the image to the size (height) of the face box itself. This ratio is treated as a height based descriptor. Figure 2.5 illustrates this. It may be noted that in cases where all individuals photographed are at the same distance from the camera, the size of their individual face boxes will almost be the same. Consequently, the distance between the face box and the base of the image can provide a rough idea of their relative heights.

Thus, 73 attribute features combined with 12 color based descriptors and 1 height descriptor together form an 86 dimensional feature vector for our future experiments.

Chapter 3

Datasets

Due to the popular implicit assumption that a detected face is independent of other detected faces in the same image, most face recognition datasets do not have images containing multiple individuals. In carefully controlled datasets, since the lighting, background and other imaging parameters are constant across photographs containing a single person, in theory, it is possible to simply assume that photographs of two different individuals belong to the same group shot. Although this can be used to generate a synthetic dataset, it would be incompatible with our basic premise of trying to learn true correlations between feature components that arise in group photographs due to similarities in lighting, pose, expression, etc. Consequently, we use two other datasets - one from a television show and another from a personal photo album.

3.1 The Buffy Dataset

First used by Everingham et al. [27], the Buffy dataset consists of frames extracted from two episodes (Season 5, Episodes 2 and 5) of the popular television series Buffy the Vampire Slayer. Each image frame contains manual annotations corresponding to automatically detected face boxes. Everingham et al. used this dataset to test their automatic character naming system for TV shows. Their system used features computed around fiducial points in a face image, clothing color based descriptors, visual speaker identification and even included speaker

information from subtitles. Around 69% accuracy was reported for recognizing all detected face images in both episodes, while accuracy was around 80% when labelling 80% of the data which had high recognition confidence.

In our case, after retaining characters that occur in group shots in both episodes, our working set consists of eight individuals. Each automatically detected face box for the eight retained characters was passed through the attribute generation pipeline described in Section 2.1, followed by the computation of new descriptors from Section 2.2.2. We use data from one full episode (Episode 2) for training and the other episode (Episode 5) for testing. We identify two subsets of feature data for each episode. The first, which we call *group-data*, consists of features computed for retained characters from group shots alone. By group shots, we mean images that contain more than one character from our working set. The other subset consists of features computed for each individual from all occurrences of the individual in an episode (not just group shots). We call this *all-data* and it includes features that are part of group-data. Figures 3.1 and 3.2 show the number of samples available for each individual in group-data and all-data for the train and test episode respectively. Tables 3.1 and 3.2 summarize the total number of frames containing each pair of individuals in the train and test episodes respectively. It may be noted that a frame that contains images of more than two individuals will be counted multiple times in Tables 3.1 and 3.2. For example, if a frame captures P_1, P_2 and P_3, then this frame can provide samples for the pairs (P_1, P_2), (P_2, P_3) and (P_3, P_1) and thus will be counted three times in such tables.

By design, our relative models will only be able to work on group-data. On the other hand, existing techniques that we implement as baselines can be trained on all-data from the training episode. Factoring these existing techniques, for a fair comparison, we use all-data for training our baseline models and group-data for training our relative models but test both models on group-data from the test episode.

There are several features of the Buffy dataset that need to noted. Since the dataset is derived from a television show, factors such as makeup and artistic camera effects are common. A large fraction of the show is shot at night (as

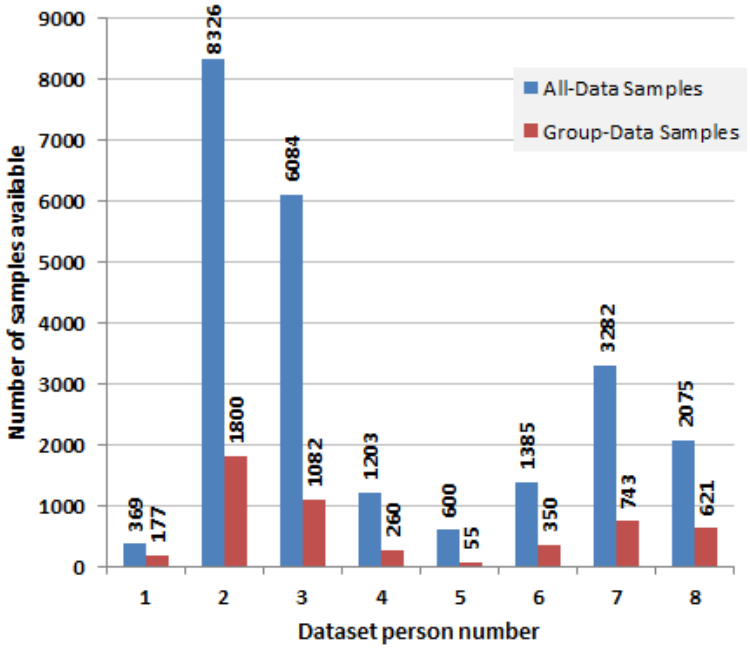

Figure 3.1: Number of samples available per individual from group shots and all shots in the training episode of the Buffy dataset

Table 3.1: Distribution of frames containing each pair of individuals in the training episode of the Buffy dataset. E.g. Person 1 and Person 3 occur together in an image 13 times whereas Person 1 and Person 5 are never seen together.

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	
	0	13	0	0	0	0	165	P_1
		597	205	36	347	696	12	P_2
			36	21	0	31	442	P_3
				0	0	80	0	P_4
					0	0	0	P_5
						0	3	P_6
							0	P_7
								P_8

expected for a show about vampires!) and so artificial directional lighting is extensively used. Since the frames are extracted from a video, a small amount of motion blur is present. In addition, frames capture a snapshot of various character activities and movements (such as talking, walking or even fighting) and in some sense, represent a wider variety of poses and expressions than what one would encounter in other datasets, including personal photo albums.

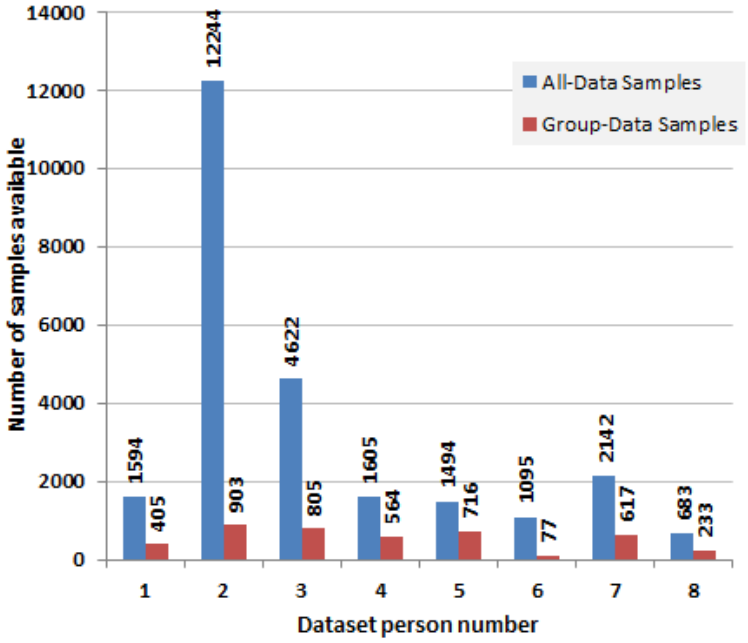

Figure 3.2: Number of samples available per individual from group shots and all shots in the test episode of the Buffy dataset

Table 3.2: Distribution of frames containing each pair of individuals in the test episode of the Buffy dataset

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	
	0	0	278	0	0	282	49	P_1
		159	269	91	2	264	184	P_2
			0	653	21	0	0	P_3
				0	0	220	0	P_4
					0	0	0	P_5
						54	0	P_6
							39	P_7
								P_8

3.2 A Personal Photo Album

One of the main arguments supporting the relative recognition model is to exploit commonalities in various imaging factors that arise naturally in group photographs. Since we were unable to find any popular labeled datasets that satisfy this requirement, we use one of our own personal photo albums. The dataset consists of approximately 1700 pictures captured on seven different days

spread over a three month period. Four different digital cameras were used. The dataset contains a mix of images captured in bright daylight, moderate indoor lighting and camera flash. Most images have individuals posing for the camera and hence contain frontal shots. This is unlike the Buffy dataset where actors are usually instructed not to look at the camera. Automatically detected face boxes were manually annotated to obtain 116 unique individuals. Figure 3.3 shows a distribution of the number of images available per person in the entire dataset.

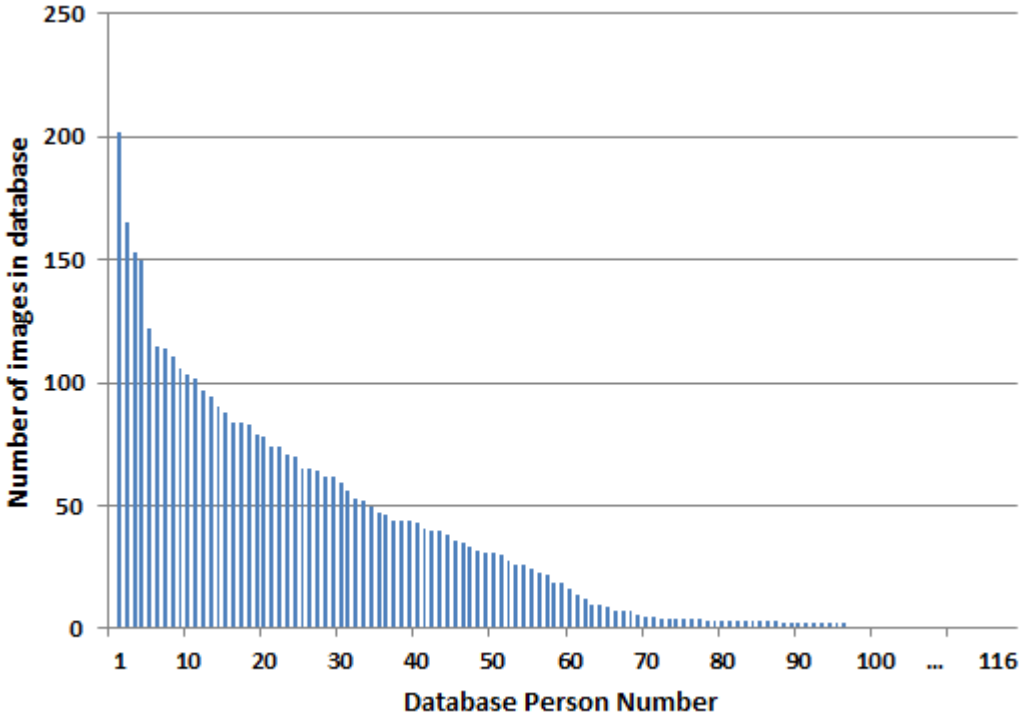

Figure 3.3: Number of images available per individual in the personal photo album dataset

We randomly select 70% of the images for training and use the rest for testing. It may be noted that the entire image file is considered as part of the training or testing process, irrespective of the actual number of individuals in the image. In order to build good relative models, each with an appreciable amount of data, we restrict all our future experiments to two sets of people. The first is a set of individuals that occur in at least 80 training images. This constitutes 6 individuals (P_1 to P_6). The second set consists of individuals that occur in at least

65 training images. This consists of 12 individuals (P_1 to P_{12}). Figure 3.4 lists the number of group-data and all-data images for each of the 12 individuals. Unlike the Buffy dataset, most photographs in this dataset are group shots. Table 3.3 shows the number of group shots for each pair of individuals. One may notice that this table is not sparse like Tables 3.1 and 3.2 for the Buffy dataset.

Although the second set is meant to contain a larger number of people and pairs than the first set, the number of training instances available is low for individuals P_7 to P_{12} and pairs involving them. With less data and more pairwise classes, we expect the performance of our relative model to deteriorate for the 12 person dataset and hence use this set to observe and understand the reduction in accuracy.

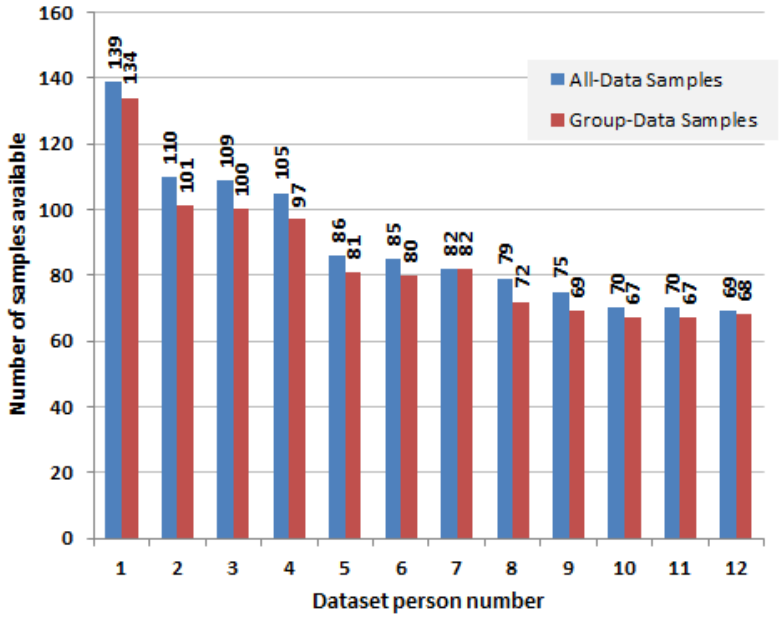

Figure 3.4: Number of samples available per retained individual from group shots and all shots in the personal photo album dataset

Table 3.3: Distribution of frames containing each pair of individuals in the personal photo album dataset

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	
	47	40	56	51	52	57	43	24	40	37	42	P_1
		53	60	45	37	44	29	53	35	38	49	P_2
			52	48	41	38	29	40	40	43	41	P_3
				47	38	31	24	39	34	33	27	P_4
					45	36	27	26	34	36	36	P_5
						44	28	16	27	32	32	P_6
							42	22	31	38	38	P_7
								23	27	31	35	P_8
									25	20	25	P_9
										31	30	P_{10}
											45	P_{11}
												P_{12}

Chapter 4

First Steps ...

Face recognition systems to date treat each detected face in a photograph independent of all other faces in the process of making a recognition decision. While this seems appropriate for photographs containing only one individual, for group photographs containing at least two individuals, there may be valuable information in other individuals' face image which can lead to a positive improvement in overall recognition accuracy. Instead, when presented with two face boxes and their corresponding feature vectors $\vec{x}_1$ and $\vec{x}_2$, a regular model (henceforth termed the baseline model) computes a recognition probability for the pair of people (P_a, P_b) using

$$P(P_a, P_b \mid \vec{x}_1, \vec{x}_2) = P(P_a \mid \vec{x}_1)\ P(P_b \mid \vec{x}_2) \tag{4.1}$$

i.e. the probability of feature vector $\vec{x}_1$ belonging to person P_a is completely independent of feature $\vec{x}_2$ belonging to person P_b although $\vec{x}_1$ and $\vec{x}_2$ are derived from the same image [1]. Is this independence assumption warranted? If not, how can we model the dependency? As a first step in answering this question and to transcend the independence assumption, we look at factors that are common to multiple individuals in a group photograph.

[1] $P(P_a \mid \vec{x}_1)$ is a posterior probability and is not measured directly. Recognition models for each individual provide $P(\vec{x}_1 \mid P_a)$. Assuming equal base probability $P(P_a) = P(P_b)\ \forall a, b$ and ignoring $P(\vec{x}_1)$, $P(\vec{x}_1 \mid P_a)$ is used in place of $P(P_a \mid \vec{x}_1)$ since it maintains the same relative order across individuals

4.1 Common Conditions in a Group Photograph

Consider Figure 4.1 which is a group photograph of six individuals. One can readily notice various similarities between the many faces. Due to a fixed light source (in this case, the sun), shadows are consistently towards the left of the photograph. All faces share the same light source color and illumination strength. Since each person is almost equally distant from the camera, the size of detected face boxes will be similar. All individuals are standing on the same ground plane and the location of their detected face boxes will be indicative of their heights in real life. As can be expected in personal photographs, there is a strong correlation in the direction of gaze (or pose) of individuals. People are usually seen either posing for the camera or gazing at a common point of interest. Further, most individuals would have the same expression (smiling, laughing, neutral, etc.). More subtle traits are effects introduced by the camera - color balance, spectral response, exposure setting, noise and even motion blur will all be similar for every face in a group photograph.

As an example of where these commonalities can be used, consider the following scenario. In an unconstrained recognition setting, when an input face box is very different from the training examples, recognition rates may not be good. This can be attributed to the fact that during training, raw values of feature components are learned. In the testing stage, these raw values may be very different - for example, under bright illumination and heavy shadow, an intensity based feature or even a texture based feature for a particular person will produce a vastly different response than what was seen for the same individual during training. On the other hand, one can expect a relative feature to perform better. Such a feature would describe the relative variation between two individuals for the same characteristic trait. For example, a skin color based relative feature would learn that Person 1 has a darker skin tone than Person 2. So, even if the measured skin tones of Person 1 and Person 2 are very different under bright illumination, the fact that Person 1 has a darker skin tone will remain true.

Correspondingly, for features that are less dependent on color - such as the mustache attribute - factors inherent to the imaging process can produce a similar

Figure 4.1: Sample group photograph from the personal photo album dataset

effect. For example, assume Person 3 has a strong mustache attribute, stronger than that for Person 4 in training images. In case of a blurry or low resolution test image, the mustache attribute generator will have decreased confidence in its decision about Person 3. Therefore, the mustache attribute value for Person 3 may not be numerically as high as the value seen during training[2]. In the test image, Person 4 would also have a blurry or low resolution face box and consequently low confidence in his/her attribute value. But one can expect the relative order of the attribute values to remain the same - i.e. mustache attribute value for Person 3 although numerically low, will still be higher than that for Person 4. It is this relative information that we hope to capture.

In addition to higher-lower comparisons, one can also attempt to learn

[2]Since attributes are decision values from SVMs, low confidence would translate to an output value closer to 0 when trained with $+1$ and -1 as labels

raw values for feature components as usual. For instance, one can learn that person 5's frowning attribute is around 0.9 and the same attribute for person 6 is around -0.4. In addition to this, a relative model would also learn how these two values vary when person 5 and person 6 occur together in an image. Higher-lower characteristics can be implicitly learned. Maybe person 5 finds person 6 annoying, but person 6 is blissfully unaware of these feelings! A good relative model would even be able to capture such subtle correlations.

In an attempt to build these relative models, we first consider the distribution of data in any normal dataset. As a starting point, we restrict ourselves to building models that can capture relative information for each *pair* of individuals in a dataset.

4.2 There is Always Less Group Data!

For the sake of this discussion, consider our training dataset to be a personal photo album comprised of P individuals $1, 2, \cdots, P$. Consequently, there would be $P(P-1)/2$ pairs of people. Let n_p denote the number of photographs in which person p occurs. A regular face recognition system, building an individual independent model for each person, would be able to use all n_p images to build a model for person p. On the other hand, a relative face recognition system would have significantly lesser data for each of its classes encompassing a pair of individuals. Our decision rule in trying to identify a pair P_1 and P_2 from their feature vectors $\vec{x}_1$ and $\vec{x}_2$ is essentially

$$\langle P_1, P_2 \rangle = \underset{\langle P_a, P_b \rangle,\ a \neq b}{\arg\max}\ P(P_a, P_b \mid \vec{x}_1, \vec{x}_2) \tag{4.2}$$

where P_a and P_b range over all P individuals.

Although there are $P(P-1)/2$ pairs of individuals, for a recognition system the order of people within a pair matters. For example, it is necessary to know that face box 1 corresponds to person x and face box 2 corresponds to person y. (x, y) in this case is an ordered pair. Thus, instead of $P(P-1)/2$ classes, one has

to deal with $P(P-1)$ classes[3]. On an average, this implies $n_p/(P-1)$ images exist per class. There may be many relative classes with very few or no images at all. All of this translates to the fact that each relative model will almost always have less data to train with than the individual independent models. In the limiting case with a large number of images, although each relative model has access to less data than each independent model, this may no longer be a practical issue.

Drawing an analogy to our group-data and all-data qualifiers from Chapter 3, group-data may at best be the same as all-data if all images are group shots. Group-data for an individual is a collection of all instances where the individual appears with at least one other person. For each relative model containing this individual and an other person, a *part* of the group-data containing the same two people can be used. Thus, each relative model will have access to part of the group-data while the baseline model will have access to all-data.

With this analysis complete, we now formulate a framework for our experiments. This framework is used to evaluate both baseline and relative models.

4.3 Experiment Framework

In order to enable easy comparison of accuracy values we formulate our baseline experiments on the same lines as those for our relative models. Specifically, our relative models will be built to recognize a pair of individuals at once. If a picture contains n individuals, then $n(n-1)$ ordered pairs are possible. Each of these is treated as a separate recognition problem. Hence, our baseline model is also presented with a pair of individuals in each trial. Since the baseline model involves an independence assumption between detected faces in the same picture, the model effectively produces two separate recognition decisions for this input pair of individuals. Mathematically, the baseline model is presented with two feature vectors $\vec{x}_1$ and $\vec{x}_2$ from two face boxes and tasked with computing $P(P_a, P_b \mid \vec{x}_1, \vec{x}_2)$.

[3]The model for class (x, y) may be highly correlated with that for (y, x). It may be possible to build one only one model - say for (x, y) - and permute the input feature vector instead to mimic the other model. Nonetheless, the system should be capable of treating (x, y) and (y, x) as two separate classes and it is this distinction that we wish to highlight.

It does so using equation 4.1 by generating two separate probability estimates $P(P_a \mid \vec{x}_1)$ and $P(P_b \mid \vec{x}_2)$.

As mentioned in Chapter 3, we present results on two datasets. For the Buffy dataset, all-data represents significantly more samples than group-data. While our relative models are forced to use group-data, the baseline models can be trained with all-data. To mimic existing recognition systems, we train our baseline models on all-data. At the same time, we also report results on baseline models trained using group-data so that one may better appreciate the change in accuracy offered by our relative models due to their design rather than lesser training data. In addition, when recognizing a particular pair of individuals, if no training samples of that pair were encountered previously, we simply omit all occurrences of this pair during testing.

In the personal photo album dataset, unlike the Buffy dataset, nearly all images are group shots with very few containing only one individual. Preliminary experiments failed to show a significant difference in accuracy between baseline models trained using group-data and all-data. For improved clarity, we present results using baseline models only trained on all-data and drop the all-data qualifier for this dataset. Further, as mentioned in Section 3.2 we present two sets of experiments on the personal photo album dataset - one consisting of 6 individuals and another 12 individuals.

For both the baseline, as well as our relative models, accuracy is computed on a *per-person* basis - i.e if a model recognizes one person correctly, but makes a mistake with the other, this is counted as 1 correct recognition and 1 wrong recognition for computing the net accuracy. In effect, the baseline models, although trained with more data, are presented with exactly the same test samples as our relative models. Thus, one can directly compare accuracy numbers generated by these models.

For all future experiments, we present results using three feature vectors - the first uses only attribute features (Section 2.1), the second only the new color and height based descriptors (Section 2.2.2), while the third uses both.

4.4 Nearest Neighbor Model

Our first baseline experiment involves building a nearest neighbor model using each individual's feature vectors. The model is simply a collection of all feature vectors seen during training along with the corresponding person label. In testing, when presented with a test feature vector, the model searches for the nearest training vector using the euclidean distance metric. The identify of this nearest vector is then assigned to the test vector. Table 4.1 presents accuracy results.

Table 4.1: Nearest neighbor baseline accuracy (in percentage)

	Buffy		**Photo Album**	
	All-Data	Group-Data	6 people	12 people
Attributes	50.79	45.45	85.96	81.06
Attributes + new descriptors	51.18	45.92	83.00	78.66
New descriptors only	33.36	30.53	26.10	20.28

From the table, we see that models trained using all-data for the Buffy dataset indeed perform better than those trained on group-data. This increase is expected due to all-data containing more samples than group-data. Also, for the personal photo album, the 12 people subset has lower accuracy than the 6 people subset. Due to an increase in the number of classes, the larger subset is expected to be harder by choice. One may also see that our new color and height based descriptors perform rather poorly on their own. Although better than random chance, they are no match for the attribute based features. This is expected since the 13 new descriptors were designed to be complementary to the 73 attribute features and not meant to work on their own. For the Buffy dataset, including the new descriptors along with attributes causes a slight increase in accuracy. One explanation for this phenomenon is that characters in the show use carefully planned makeup and lighting that make them appear the same across episodes. Consequently, our color based descriptors are able to learn this and perform well. On the other hand, there is a decrease in accuracy when the new descriptors are included with attributes for our personal photo album dataset. Since the new descriptors are largely based on characteristics that change drastically across

images - such a decrease in accuracy may be expected. We hope to exploit the information encoded in the new descriptors through our relative models later on.

4.5 Baseline Gaussian Model

An alternative to using nearest neighbor classification is to model the class conditional density directly. One such model is a Gaussian represented as

$$P(P_a \mid \vec{x}_1) = \mathcal{N}(\mu_a, \Sigma_a) \tag{4.3}$$

where μ_a is the average feature vector for P_a and Σ_a is the covariance for these vectors - both computed from training examples. While we are able to estimate a full (86×86) covariance matrix for the Buffy dataset in the all-data case, we are unable to do so for other datasets which are relatively smaller in size. These include Buffy group-data and the photo album dataset. We believe that our photo album dataset is representative of the size of normal photo albums which in turn are moderately sized. Hence, this problem may persist in practice. Instead of the full covariance matrix in these cases, we learn a diagonal covariance matrix - effectively learning each feature component as an independent one-dimensional Gaussian. Results from this experiment can be found in Table 4.2.

Table 4.2: Gaussian baseline accuracy (in percentage)

	Buffy		**Photo Album**	
	All-Data	Group-Data	6 people	12 people
Attributes only	56.18	46.11	87.43	79.86
Attributes + new descriptors	53.30	43.17	86.69	79.64
New descriptors only	32.15	13.16	29.06	13.02

Except for the 12 people photo album, the Gaussian model outperforms the nearest neighbor model using attributes or attributes in combination with new descriptors. In all 4 data partitions here, introducing our new descriptors consistently drops recognition accuracy when compared to just using attribute features.

Chapter 5

Building Relative Models

Our main goal in this work is to build relative face recognition models that can benefit from naturally occurring commonalities in the face images of two or more individuals from the same group photograph. Towards this objective, we defined additional color and height based descriptors that can encode common traits as part of our feature vector. We now explore various techniques of building relative recognition models to harness this encoded information. In general, one can build two types of probabilistic models - a conditional model and a joint model.

5.1 Building a Conditional Model

In light of the data scarcity issue (Section 4.2), one would like to retain the more reliable individual independent model, but still add on the usefulness of a relative model. One way to do this is to formulate the relative model as a conditional probability model. Let the feature vectors corresponding to face boxes 1 and 2 be $\vec{x}_1$ and $\vec{x}_2$ respectively. The recognition framework for identifying two people P_a and P_b would be

$$P(P_a, P_b \mid \vec{x}_1, \vec{x}_2) = P(P_a \mid \vec{x}_1)\ P(P_b \mid \vec{x}_1, \vec{x}_2, P_a) \quad (5.1)$$

$$= P(P_b \mid \vec{x}_2)\ P(P_a \mid \vec{x}_2, \vec{x}_1, P_b) \quad (5.2)$$

Note that equation 5.1 and equation 5.2 although mathematically the same, may produce different results. Various ad hoc techniques can be used to combine

the two estimates, including an arithmetic mean or a geometric mean.

Any suitable model can be used to estimate $P(P_a \mid \vec{x}_1)$ or $P(P_b \mid \vec{x}_2)$ e.g. the Gaussian model described in Section 4.5. Unfortunately the conditional probabilities $P(P_a \mid \vec{x}_2, \vec{x}_1, P_b)$ and $P(P_b \mid \vec{x}_1, \vec{x}_2, P_a)$ are both very hard to estimate owing to the fact that both $\vec{x}_1$ as well as $\vec{x}_2$ are real valued vectors and that training data is already scarce. Hence, it is tough to obtain even an approximate estimate which can be used for testing.

5.2 Binary Relative Conditional Model

To bypass this problem with real valued vectors, we make a simplifying assumption and define the conditional in terms of a relative binary feature vector. For two input feature vectors $\vec{x}_a$ and $\vec{x}_b$ from two face boxes in the same image, let $\vec{C}_{ab}$ denote the relative binary feature vector of $\vec{x}_b$ with respect to $\vec{x}_a$ defined as

$$C^i_{ab} = \begin{cases} 1 & \text{if } x^i_a \geq x^i_b \\ 0 & \text{if } x^i_a < x^i_b \end{cases} \tag{5.3}$$

where i ranges from 1 to the total number of dimensions D of feature vectors $\vec{x}_a$ or $\vec{x}_b$.

The training phase involves learning the probability of feature component i being greater for person a when compared to the same feature component for person b. This training uses images where person a and person b occur in the *same* photograph. Mathematically, setting $z_1 = \text{count}(x^i_a \geq x^i_b)$ and $z_0 = \text{count}(x^i_a < x^i_b)$

$$P(C^i_{ab} = 1) = \frac{z_1}{z_0 + z_1} \tag{5.4}$$

$$P(C^i_{ab} = 0) = \frac{z_0}{z_0 + z_1} \tag{5.5}$$

The drawback of such a counting estimate is that the probability $P(C^i_{ab})$ can be perfectly zero if every pair of feature components compared bear the same relationship. In practice, while a low probability value is acceptable, a perfect zero can cause instability in decisions. Moreover, the zero probability can be seen as a byproduct of having to work with limited data. As a quick-fix to this problem, we

disturb each probability estimate $P(C^i_{ab})$ by a small amount using

$$P(C^i_{ab}) = P(C^i_{ab}) + \frac{\text{sign}(0.5 - P(C^i_{ab}))}{z_0 + z_1 + 1} \tag{5.6}$$

This changes the probability by a fraction $1/(z_0 + z_1 + 1)$ in such a way that values less than 0.5 are increased and values greater than 0.5 are decreased - moving both away from zero and one respectively. Intuitively, for each probability estimate $P(C^i_{ab})$ we assume one additional data point that belongs to the category with lower probability. While this technique successfully eliminates perfect zero probabilities, it does so in a controlled manner. If the probability estimates already counted a large number of data points, disturbing with one additional point will not make much difference. On the other hand, when estimates are based on very few samples, confidence on these estimates are also low. In such cases, adding one additional point may effect a substantial but warranted change towards probability 0.5.

The conditional probability for the testing phase can now be approximated in terms of this new metric as

$$P(P_b \mid \vec{x}_1, \vec{x}_2, P_a) \approx \prod_{i=1}^{i=D} P(C^i_{ab} = C^i_{12}) \tag{5.7}$$

This involves assuming individual components of the relative binary feature vector are independent of each other. The value for $P(P_a \mid \vec{x}_2, \vec{x}_1, P_b)$ can be computed similarly.

Intuitively, the learning phase in equations 5.4 and 5.5 learn the probability that a particular feature component is numerically higher for one person when compared to another. For example, if person a is male and person b female, then the male attribute for person a will usually be higher than the male attribute for person b. Due to various factors, this inequality may not always hold, but one would expect the probability $P(C^{male}_{ab} = 1)$ to be close to 1. Correspondingly, $P(C^{male}_{ab} = 0)$ would be close to 0. In case an attribute is chosen where the two individuals being compared have nearly similar values, then both these probabilities would be close to 0.5.

It may be noted that the above conditional approximation is used along with the individual independent models following equations 5.1 or 5.1 or both. For our

experiments, the baseline estimate is provided by Gaussians described in Section 4.5. From our previous experiment, we know the baseline Gaussian using attributes alone performs best and so, it is this model that we use. As shown in equations 5.1 and 5.2, we compute two estimates of the same quantity $P(P_a, P_b \mid \vec{x}_1, \vec{x}_2)$ using baseline models for P_a and P_b. We combine these two estimates using ad hoc fusion techniques such as arithmetic mean and geometric mean. It is important to note that the actual raw values of feature components are completely ignored by the binary conditional approximation model. Results for this technique are presented in Table 5.1.

A first glance at this table reveals the geometric mean fusion to be better than arithmetic mean fusion. Although this is not true for the Buffy dataset when using our new descriptors alone, the difference between these accuracies is small compared to the overall boost from the baseline. We also notice that the geometric mean fusion consistently provides higher accuracy than the baseline technique. The increase is rather modest around 0.4% for the 6 people photo album and 2.5% for the 12 people photo album. For the Buffy dataset, it is 8%.

Thus, we see an increase in accuracy when using baseline models trained on attribute data and binary conditional models trained on our new descriptors. Although the new descriptors are rather weak as baseline feature vectors, using them in the conditional model does help. This shows the presence of valuable information in color and height based descriptors that have long been ignored in mainstream face recognition. An interesting observation is that attributes along with the new descriptors do not work as well in the conditional model. Attribute values are generated by SVMs, and although we consider the raw SVM output to be an indication of the degree of presence or absence of an attribute, the SVM is optimized for a different purpose - namely maximum margin separation. We conjecture that this is the reason for poor performance of the conditional model when using both attributes and our new descriptors.

As a slightly more rigorous demonstration of the accuracy increase provided by the conditional model, we perform another experiment. A baseline model is trained with a random subset of attribute features. The number of attributes

Table 5.1: Binary conditional model accuracy (in percentage)

	Buffy		**Photo Album**			
			6 people		12 people	
Conditional model features	Arith. mean	Geom. mean	Arith. mean	Geom. mean	Arith. mean	Geom. mean
Attributes only	55.93	59.17	83.25	84.48	78.44	79.71
Attributes + new descriptors	59.25	62.59	83.99	85.96	79.04	81.28
New descriptors only	**64.63**	64.22	79.92	**87.80**	66.42	**82.33**
Baseline (Attributes only)	56.18		87.43		79.86	

per subset is varied from zero to all available attributes. For each attribute subset, a baseline recognition experiment is performed as detailed previously. Correspondingly, with each baseline model thus trained, a binary conditional model trained on all of our new height and color descriptors is used to provide a relative decision. It may be noted that the baseline model uses a different number of attributes on each trial, while the binary conditional model always uses all of our new descriptors. The entire experiment is repeated 30 times and results from each run are averaged. Figure 5.1 shows the output from this experiment for the Buffy dataset. Figure 5.2 shows results for the personal photo album with 6 people. As can be seen in both figures, the conditional model using our new descriptors consistently provides higher recognition accuracy than an existing baseline model. While the difference in accuracy using a low number of attribute features is rather high, as the number of attributes increase, the improvement provided by color descriptors diminishes as expected. Nonetheless, even when all available attributes are used, the conditional model still outperforms the baseline model. Specifically for the Buffy dataset, our conditional model (using 13 color/height descriptor components) along with two or three attribute components provides higher accuracy than baseline models using all 75 attributes.

5.3 Building a Joint Model

Although it performs well, our binary approximation of the conditional model completely ignores the raw value of feature components - instead looking only at the relative higher-lower relationship between the two components. While still working in the realm of a relative model, a joint probability model can learn raw feature values for two individuals' face regions as well as a correlation between them for each feature component. As a first step, we build a nearest neighbor model to encode joint information. For each joint model presented here, group-data is used to train the model for both the Buffy dataset and the personal photo album.

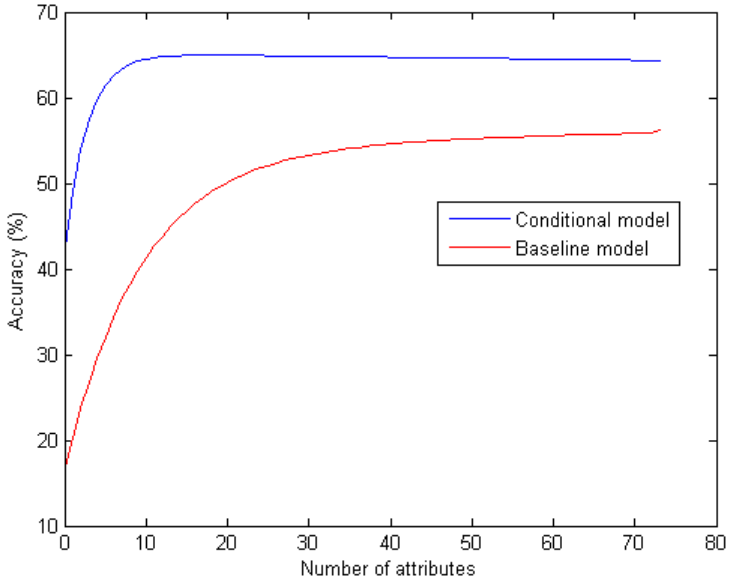

Figure 5.1: Variation in accuracy with number of attributes used for baseline and conditional models on the Buffy dataset

5.4 Nearest Neighbor Joint Model

Along the same lines as our baseline nearest neighbor model (Section 4.4), we build a nearest neighbor model to learn the joint information between the detected face box of two different individuals. The two feature vectors extracted from each face box are simply concatenated together resulting in one long feature vector which is then fed to the model. In training, the model simply retains all such concatenated vectors for each ordered pair of individuals. Thus, with P individuals, we learn $P(P-1)$ pairwise classes. With $\vec{x}_1$ and $\vec{x}_2$ as the feature vectors extracted from each face box, if the class (P_a, P_b) is trained with $(\vec{x}_1, \vec{x}_2)$, then the class (P_b, P_a) is trained with $(\vec{x}_2, \vec{x}_1)$. For example, when training with just the 73 attributes, $(\vec{x}_1, \vec{x}_2)$ is a $2 \times 73 = 146$ element vector. For testing, when presented with a new concatenated feature vector, the model simply finds the nearest feature vector in its training set using a euclidean distance metric. If this nearest feature vector belongs to model (P_c, P_d), then the first detected face box is assigned P_c and the second P_d. Results from this experiment are shown in Table 5.2.

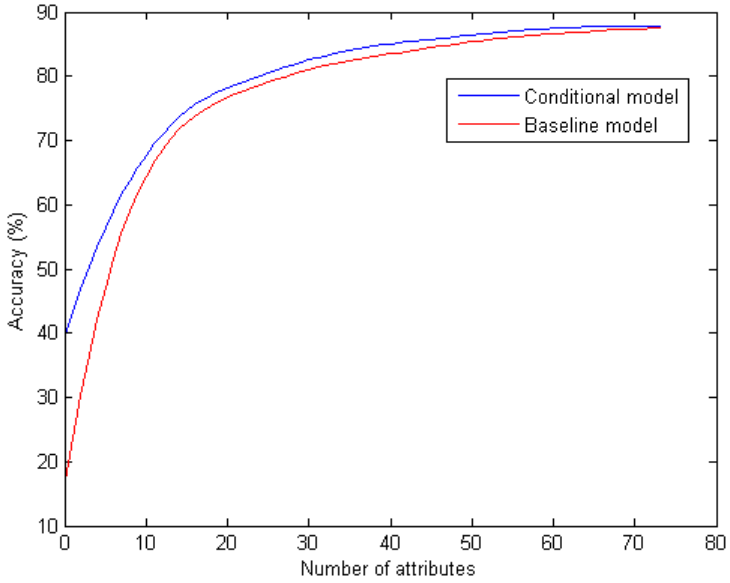

Figure 5.2: Variation in accuracy with number of attributes used for baseline and conditional models on the 6 people photo album dataset

Table 5.2: Nearest neighbor joint model accuracy (in percentage)

	Buffy	**Photo Album**	
		6 people	12 people
Attributes only	40.82	80.78	74.70
Attributes + new descriptors	40.85	78.07	70.73
New descriptors only	38.76	34.72	20.88

Compared to our nearest neighbor baseline models from Table 4.1, we notice a decrease in accuracy in all cases except when using our new descriptors on their own. For the Buffy dataset, one would have to compare these numbers to the Buffy group-data baseline models that have access to the same training examples. On an average, when using attributes alone or in combination with the new descriptors, accuracy on the Buffy dataset and the 6 people photo album decreases 5%, while the 12 people photo album experiences a decrease of $6\% - 8\%$. Before proposing an explanation for the decrease in accuracy when using attributes and the increase when using our new descriptors alone, we also build a Gaussian joint Model.

5.5 Gaussian Joint Model

Considering that for baseline experiments, the Gaussian model performed better than the nearest neighbor model, here too, we try to learn a single Gaussian for each pair of people. The input vector for such a model would again simply be $(\vec{x}_1, \vec{x}_2)$ a concatenated version of feature vectors for two individuals (P_a, P_b). As before, we would have to learn $P(P-1)$ pairwise models for P individuals in the dataset. Unfortunately, due to data scarcity, it was observed that the full covariance matrix for the pairwise Gaussian is inaccurate (a similar effect was described in Section 4.5 when working with group-data for both Buffy and photo album dataset). This is reflected in the very low recognition rates and a singular covariance matrix.

It is important to note that this is indeed a very real problem especially when working with joint models using personal photo albums as datasets. For example, although a photo album may contain around 1000 images, even with 10 people there are 45 unordered pairs. Each of our feature vectors for one face box consists of 86 components. The Gaussians are trained on two such feature vectors concatenated together - totaling 172 components. For a non-singular estimate of the full covariance matrix, at least 172 good photographs need to be present for every pair of individuals - which is quite rare. While it is true that group photographs containing a large fraction of the 10 people would help in satisfying this requirement, it is often seen that in case of personal photo albums, the chances of finding multiple group shots containing a large number of individuals are often low. Even if this occurs, in many cases the photographs are highly correlated themselves (taken a few seconds apart) and usually carry redundant information which does not help in building a good estimate for the full covariance matrix.

To circumvent this pitfall, we take a closer look at our joint model. Assuming independence between feature vector components for an individual, we only seek to learn correlations between the same feature component for every pair of people. This is similar in spirit to our approximate conditional models that only compare corresponding feature components. Thus, we propose a simpler joint Gaussian model which learns multiple 2 dimensional Gaussians for each feature

component, for each pair of individuals. Mathematically,

$$P(P_a, P_b \mid \vec{x}_1, \vec{x}_2) = \prod_{i=1}^{i=D} \mathcal{N}(\vec{\mu}_i^{ab}, \Sigma_i^{ab}) \tag{5.8}$$

where D is again the total number of feature vector components, $\vec{\mu}_i^{ab}$ is the 2 dimensional mean for the concatenated vectors $[x_a^i \; x_b^i]$ seen during training, Σ_i^{ab} is the 2×2 full covariance matrix for the same vectors and $\mathcal{N}(\vec{\mu}_i^{ab}, \Sigma_i^{ab})$ is the single Gaussian learned for feature component i for the pair of individuals P_a and P_b. Thus, with a dataset consisting of P people, we would learn $DP(P-1)$ 2-dimensional Gaussians.

A minor issue when combining multiple low probability values as a product is one of floating point underflow. To mitigate this, we work in the log probability space wherever possible. i.e.

$$\log P(P_a, P_b \mid \vec{x}_1, \vec{x}_2) = \sum_{i=1}^{i=D} \mathcal{N}(\vec{\mu}_i^{ab}, \Sigma_i^{ab}) \tag{5.9}$$

Results for this experiment can be found in Table 5.3.

Table 5.3: Gaussian joint model accuracy (in percentage)

	Buffy	**Photo Album**	
		6 people	12 people
Attributes only	51.21	85.22	79.26
Attributes + new descriptors	49.81	84.97	78.44
New descriptors only	36.88	43.84	25.52

Comparing this to the accuracy Table 4.2 we make a few observations. When using attributes alone or in combination with our new descriptors, net accuracy for all datasets drop when compared to corresponding baseline versions. For the Buffy dataset (compared to baseline all-data), there is a $3\% - 5\%$ decrease, and for the photo album, a $1\% - 2\%$ decrease. On the other hand, when our new descriptors are used without attributes, accuracy increases by $12\% - 14\%$ for all datasets. For the Buffy dataset (compared to baseline group-data) and the 12 person photo album, this is double the previous accuracy. Although the base accuracy is numerically low, we consider the substantial change in accuracy as a

strong indication that our joint model is able to exploit relative data encoded in our new descriptors.

To further understand the performance boost provided by our new descriptors in a joint setting, we compare the accuracy of a baseline model and a joint model both trained with just one of our new descriptors. Figure 5.3 shows results for this experiment on the Buffy dataset. Figure 5.4 shows similar results on the 6 people personal photo album. In all cases, the height based descriptor is a single number, whereas the other color descriptors each consist of 3 components - medians of hue, saturation and value computed for specific regions. In line with our previous observation, while the overall accuracy of each descriptor is numerically low, every descriptor provides a boost in accuracy when used in a joint Gaussian model.

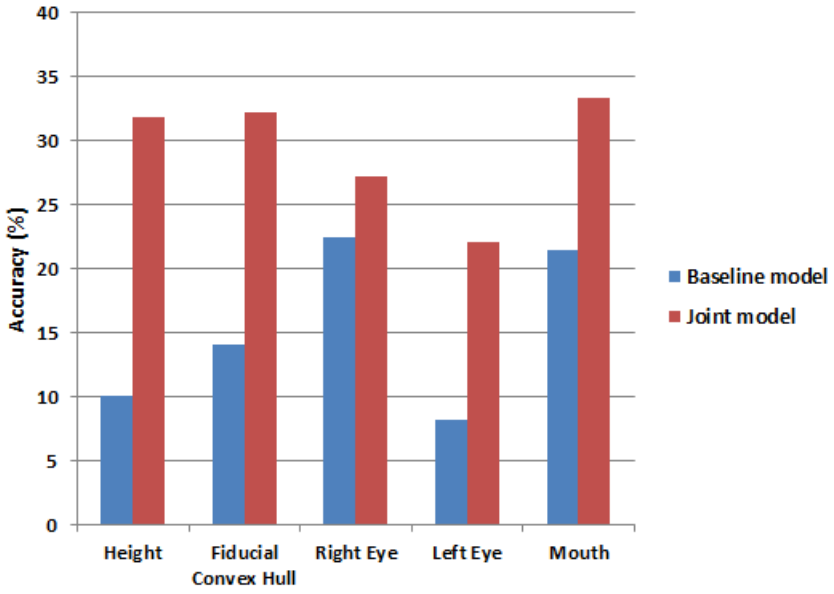

Figure 5.3: Comparison of accuracies provided by baseline and joint models using our new descriptors on the Buffy dataset

Analyzing these results, we hypothesize that the reduced amount of data available per joint model is indeed detrimental to recognition accuracy. In addition, as noted in Section 4.2, this detrimental effect due to data scarcity will persist in any moderately sized database. Nonetheless, there is an appreciable increase in

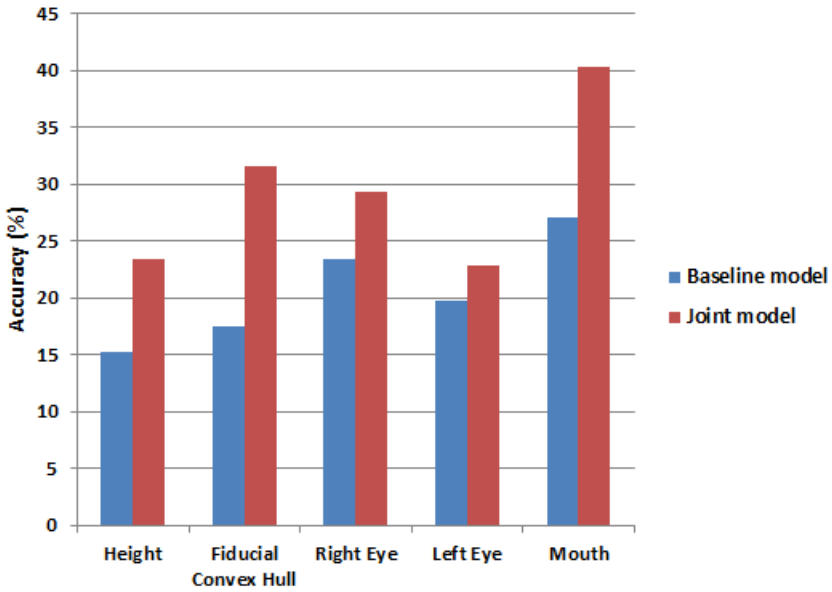

Figure 5.4: Comparison of accuracies provided by baseline and joint models using our new descriptors on the 6 people photo album

accuracy using our new descriptors alone. Also, each joint model, despite having access to few training examples, has performance that is comparable to well trained baseline techniques. Due to both these factors, we continue to explore techniques in the next chapter that can boost baseline recognition accuracy using our joint model.

Chapter 6

Baseline-Joint Fusion Techniques

For a particular pair of individuals, the joint model is trained only using images that contain both individuals. It is common for each individual from this pair to appear in many other photographs - alone or with others. Unfortunately, by its very design, the joint model is unable to utilize this data while making a decision for the particular pair of individuals considered. But when it comes to leveraging common traits in group photographs by using our new color and height based descriptors, joint models seem promising - sometimes even doubling the accuracy of baseline techniques.

Ideally, one would like to combine the baseline and the joint models. The baseline with its relatively ample training data can provide a good per-person estimate. The joint can provide improvements in case of group photographs by leveraging color and height descriptors. In fact this is exactly what the conditional model is able to do following basic laws of probability. On the other hand, strict mathematical relations that combine the joint and baseline models for making decisions are not straightforward. Mathematically, given feature vectors $\vec{x}_1$ and $\vec{x}_2$ from two face boxes in the same image, our goal is to compute $P(P_a, P_b \mid \vec{x}_1, \vec{x}_2)$. By its very design, the joint model directly outputs this probability estimate. The baseline model computes the same quantity as $P(P_a \mid \vec{x}_1)\ P(P_b \mid \vec{x}_2)$. Both of these represent the same quantity but under different independence assumptions, and so it is unclear how to properly combine the joint and baseline models. Consequently, we propose a few new techniques of our own to fuse the probability outputs from

the baseline and joint models.

For all these experiments, we use the Gaussian model as our baseline and train it on attribute features alone. As seen in Table 4.2, the highest accuracy was obtained when working with such a setup. Also, by restricting our baseline to a state-of-the-art feature set, we allow for our fusion techniques to be easily extended to existing systems that ignore color and height based descriptors.

6.1 Arithmetic and Geometric Means

Borrowing our previous notation, while the joint model explicitly computes $P(P_a, P_b \mid \vec{x}_1, \vec{x}_2)$, the baseline model simply assumes that $P(P_a, P_b \mid \vec{x}_1, \vec{x}_2) = P(P_a \mid \vec{x}_1)\ P(P_b \mid \vec{x}_2)$. Treating these as two estimates of the same quantity, we simply take their arithmetic mean and use this new quantity for decision making. Table 6.1 shows results obtained using this technique. As mentioned, the baseline technique is trained on attributes alone whereas the joint model uses the three sets of features indicated. Also included are results obtained by taking the geometric mean of the two quantities instead. Estimates from the joint and baseline models were separately scaled to lie in $[0, 1]$ before applying either technique. To do so, estimates for all joint models were divided by the highest probability value among them. A similar approach was used for normalizing the baseline model estimates.

One can make various observations from the table. Joint models built on attributes in conjunction with our new descriptors are able to outperform baseline models even with ad hoc geometric mean fusion. But there doesn't seem to be a consensus on one method that consistently performs well across all datasets. For example, Geometric mean with attributes alone (for both baseline and joint models) provides the highest accuracy for the Buffy dataset and the 12 people photo album, but not the 6 people dataset. Before reading further into these numbers, we present one final experiment involving a more structured approach to fusing the joint and baseline decisions.

Table 6.1: Combined baseline and joint accuracy (in percentage)

		Baseline	**Joint**	**Combined**	
				Arith. mean	Geom. mean
Buffy	Attributes only	56.16	51.21	56.68	**58.56**
	Attributes + new descriptors	56.16	49.81	55.96	56.90
	New descriptors only	56.16	36.88	52.72	56.37
Photo Album (6 people)	Attributes only	87.43	85.22	85.46	85.96
	Attributes + new descriptors	87.43	84.97	87.68	**87.93**
	New descriptors only	87.43	43.84	86.20	84.72
Photo Album (12 people)	Attributes only	79.86	79.26	80.76	**81.28**
	Attributes + new descriptors	79.86	78.44	79.86	80.83
	New descriptors only	79.86	25.52	78.21	78.36

6.2 Confidence Gating Technique

Since a simple technique like the geometric mean is able to provide an increase in recognition accuracy, we explore more structured approaches hoping to get a further boost by combing the joint and baseline decisions. Here, we present a technique based on confidence thresholds. Given a pair of face boxes to recognize, the attribute trained baseline model is used to obtain two probability estimates for all individuals corresponding to each of the face boxes. For each face box, we define a confidence metric as the ratio of the highest probability to the next highest probability in the list of probabilities across all individuals for this face box. Intuitively, a high value for this confidence metric would imply that the system is not confused and is relatively sure that the face box belongs to one particular individual. On the other hand, a low value for this confidence metric implies that the system thinks at least two individuals are almost equally likely candidates for this face box. The overall technique can be described using algorithm 6.1. P_1 and P_2 are the final recognition labels that our algorithm assigns corresponding to feature vectors $\vec{x}_1$ and $\vec{x}_2$ respectively.

In summary, the algorithm above uses the baseline probability for recognition if it has high confidence. If one of the two probabilities among the pair has low confidence, then the joint probability is used after fixing an identity using the other more confident probability estimate. In case both baseline estimates have insufficient confidence, then recognition is based purely on the joint estimate.

The confidence threshold is selected after a linear sweep performed through 3-fold cross-validation on our training set. Results for this experiment are presented in Table 6.2. To demonstrate the usefulness of our confidence metric, we use the same confidence threshold and run a control experiment that computes the accuracy of only those baseline test-pairs for which we have high confidence in both individuals. When compared to the regular baseline experiments, these experiments showed an increase in accuracy of $4\% - 8\%$. This indicates that our confidence metric while not perfect, can identify baseline recognition pairs that are more likely to be correct and use the joint model for others.

From the table, we see that for the personal photo album dataset with 6 and

Algorithm 6.1 Confidence Gating

Let the dataset contain P individuals $1, 2, \ldots, P$
Compute baselines $P(P_a = i \mid \vec{x}_1)$ and $P(P_b = i \mid \vec{x}_2)$ for $i = 1, 2, \ldots P$
Compute joint $P(P_a = i, P_b = j \mid \vec{x}_1, \vec{x}_2)$ for $i, j = 1, 2, \ldots, P;\ i \neq j$
$C_a \leftarrow$ Confidence metric for $P(P_a \mid \vec{x}_1)$
$C_b \leftarrow$ Confidence metric for $P(P_b \mid \vec{x}_2)$
$C_t \leftarrow$ Confidence threshold
case $C_a \geq C_t$ AND $C_b \geq C_t$
    $P_1 \leftarrow \arg\max_i P(P_a = i \mid \vec{x}_1)$
    $P_2 \leftarrow \arg\max_j P(P_b = j \mid \vec{x}_2)$
case $C_a \geq C_t$ AND $C_b < C_t$
    $P_1 \leftarrow \arg\max_i P(P_a = i \mid \vec{x}_1)$
    $P_2 \leftarrow \arg\max_j P(P_a = P_1, P_b = j \mid \vec{x}_1, \vec{x}_2)$
case $C_a < C_t$ AND $C_b \geq C_t$
    $P_2 \leftarrow \arg\max_j P(P_b = j \mid \vec{x}_2)$
    $P_1 \leftarrow \arg\max_i P(P_a = i, P_b = P_2 \mid \vec{x}_1, \vec{x}_2)$
case $C_a < C_t$ AND $C_b < C_t$
    $(P_1, P_2) \leftarrow \arg\max_{(i,j)} P(P_a = i, P_b = j \mid \vec{x}_1, \vec{x}_2)$

12 people, joint models trained on attributes and our new descriptors provide the best improvement over baseline accuracy. Thus, our joint models are successfully able to extract additional correlations from attributes and our new color and height based descriptors. Further, our confidence gating technique is able to effectively combine the joint and baseline decisions into a final better decision. On the Buffy dataset, the best joint model uses only attributes. As mentioned previously, even when working with attributes, the joint model can learn implicit correlations that arise due to common lighting, camera response, etc.

6.3 Summary of Results

Table 6.3 presents a consolidated view of the best accuracies attained and corresponding features used for each of the baseline, conditional and joint models.

Table 6.2: Confidence gating accuracy (in percentage)

		Baseline	Joint	Confidence Gating
Buffy	Attributes only	56.16	51.21	**57.21**
	Attributes + new descriptors	56.16	49.81	56.66
	New descriptors only	56.16	36.88	56.18
Photo Album (6 people)	Attributes only	87.43	85.22	87.68
	Attributes + new descriptors	87.43	84.97	**88.42**
	New descriptors only	87.43	43.84	88.17
Photo Album (12 people)	Attributes only	79.86	79.26	81.13
	Attributes + new descriptors	79.86	78.44	**81.66**
	New descriptors only	79.86	25.5	79.86

When baseline models are used along with the conditional and joint models, they are trained using the best performing feature for independent recognition - which happens to be attributes alone.

Table 6.3: Summary of accuracy using various techniques (in percentage)

	Model	**Accuracy**	**Features used**	
			Attributes	New Descriptors
Buffy	Baseline	56.18	✓	
	Conditional	64.63		✓
	Fused joint	57.21	✓	
Photo Album (6 people)	Baseline	87.43	✓	
	Conditional	87.80		✓
	Fused joint	88.42	✓	✓
Photo Album (12 people)	Baseline	79.86	✓	
	Conditional	82.33		✓
	Fused joint	81.66	✓	✓

Chapter 7

Conclusions

Face recognition systems have always built models for each individual in isolation. When presented with a group photograph, such systems assume statistical independence between detected faces. We observe various naturally occurring commonalities between face regions in the same image - all of which seem to invalidate this assumption. These commonalities include lighting, ground planes, direction of shadow and gaze etc. We take a first step in transcending the independence assumption and encode the commonalities using a few new color and height based descriptors of our own. The variations that our descriptors capture have largely been considered hindrances to face recognition and completely ignored by many systems. In addition to the new descriptors, due to the unconstrained nature of natural images, we also use state-of-the-art attribute based features.

We propose two models that recognize pairs of individuals at once from a group shot. The first conditional probability model is based on higher-lower comparisons of feature components and it uses probability outputs from existing baseline techniques during its decision making. This makes it feasible to be added on as a module in existing systems - providing a boost in recognition accuracy in case of group shots. The second model that we build captures joint probability information and models correlation across the same feature components for two people.

Evaluating performance, we see that our new descriptors generally cause a decrease in recognition accuracy when combined with attribute features used in a

baseline model. Thus, simple existing systems that assume statistical independence between faces are unable to use the new color and height information. On the other hand, the conditional probability model is able to exploit information in the new descriptors and consistently produces a boost in recognition accuracy. Although the joint probability model encounters an increase in accuracy when using the new descriptors, it takes a hit due to data scarcity - which causes the overall accuracy to be less than baseline techniques. We believe that with increasingly cheap storage and image capture technologies, a day may not be far when the data scarcity problem disappears. But for the time being, we show that fusing the joint decision with baseline decisions using a confidence gating technique can be effective.

Chapter 8

Future Work

Future work in this direction can be geared towards a few broad areas. First, it may be possible to formulate additional descriptors that can bring out various similar traits that occur naturally in group photographs. On a tangential note, such descriptors need not be restricted to the face region alone. It may be possible to generate clothing and body descriptors in a similar fashion.

Another avenue is to make the relative models better. Data from group photographs is usually scarce and so better techniques that squeeze more information from existing data may be warranted. While we consider simple nearest neighbor and Gaussian models, one may attain higher accuracies using multi-class discriminative classifiers. Also, it may be possible to combine joint estimate with the baseline in better ways.

Attacking the problem of data scarcity head-on would include building systems that can learn from a single training image for each individual or a single image for a group of individuals.

It may be possible to infer relative information by transitivity. For example, although two people may never be seen together in group shots, if each of them has been photographed with a common third person, then it may be possible to transitively infer relationships between feature components of these two people. As an extension to this, it may be possible to use transitivity relations through multiple individuals - although one would expect such inferred relations to get worse with longer transitive chains.

Finally, while we deal with pairs of people, it may be possible to use estimates from all pairs in an image to form an overall recognition decision for each person. Also, instead of pairwise models, models describing three of more individuals may be possible. Our initial calculations show an exponential increase in the complexity of this problem with increasing group sizes. But perhaps a simpler consensus strategy is possible.

Bibliography

[1] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical Report 2010-66, Microsoft Research, Redmond, June 2010.

[2] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[3] M. Turk and A. Pentland. Face fecognition using eigenfaces. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[4] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.

[5] A. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2d and 3d face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885 – 1906, 2007.

[6] X. Zhang and Y. Gao. Face recognition across pose: A review. *Pattern Recognition*, 42(11):2876 – 2896, 2009.

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

[8] A. Das, O. Manyam, and M. Tapaswi. Multi-feature audio-visual person recognition. In *Proc. IEEE Workshop on Machine Learning and Signal Processing*, pages 227–232, 2008.

[9] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[10] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *Proc. IEEE International Conference on Computer Vision*, pages 365–372, 2009.

[11] M. Guillaumin, J. Verbeek, and C. Schmid. Attribute and simile classifiers for face verification. In *Proc. IEEE International Conference on Computer Vision*, pages 365–372, 2009.

[12] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2707–2714, 2010.

[13] L. Zhang, L. Chen, M. Li, and H. Zhang. Automated annotation of human faces in family albums. In *ACM Multimedia Conference*, pages 355–358, 2003.

[14] M. Naaman, R. Yeh, H. Garcia-molina, and A. Paepcke. Leveraging context to resolve identity in photo albums. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 178–187, 2005.

[15] D. Anguelov, K. Lee, S. Gokturk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1 –7, 2007.

[16] J. Sivic, C. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *Proceedings of the British Machine Vision Conference*, 2006.

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

[18] Z. Stone, T. Zickler, and T. Darrell. Toward large-scale face recognition using social network context. *Proc. of the IEEE*, 98(8):1408 –1415, August 2010.

[19] A. Gallagher and T. Chen. Estimating age, gender and identity using first name priors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[20] A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 256–263, 2009.

[21] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35:399–458, 2003.

[22] J. Yang and C. Liu. Color image discriminant models and algorithms for face recognition. *IEEE Trans. on Neural Networks*, 19(12):2088 –2098, December 2008.

[23] C. Wang, B. Yin, X. Bai, and Y. Sun. Color face recognition based on 2dpca. In *Intl. Conf. on Pattern Recognition*, pages 1 –4, December 2008.

[24] F. Hajati, K. Faez, and S. Pakazad. An efficient method for face localization and recognition in color images. In *IEEE Intl. Conf. on Systems, Man and Cybernetics*, volume 5, pages 4214 –4219, October 2006.

[25] J. Choi, Y. Ro, and K. Plataniotis. Color face recognition for degraded face images. *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(5):1217 –1230, October 2009.

[26] E. Marszalec, B. Martinkauppi, M. Soriano, and M. Pietikinen. A physics-based face database for color research. *Journal of Electronic Imaging*, 9(1):32–38, 2000.

[27] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *British Machine Vision Conference*, 2006.